Tesis de Maestría

Descubrimiento de sitios regulatorios potenciales en las regiones intergénicas de Mycobacterium tuberculosis utilizando técnicas de minería de datos

Henrión, Guillermo Gabriel

2013

UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Descubrimiento de sitios regulatorios potenciales en las regiones intergénicas de *Mycobacterium tuberculosis* utilizando técnicas de minería de datos

Tesis presentada para optar al título de Magister de la Universidad de Buenos Aires en Explotación de Datos y Descubrimiento del Conocimiento

Autor: Lic. Guillermo G. Henrión

Director: Dr. Marcelo A. Soria

Co-Directora: Dra. Ana S. Haedo

Buenos Aires, 2013

Contenido

Resumen

El objetivo del presente trabajo es descubrir potenciales sitios regulatorios dentro de las regiones intergénicas de *Mycobacterium tuberculosis*. Se utilizaron experimentos con microarreglos (microarrays) depositados en el NCBI, a los cuales se le aplicaron diversos algoritmos de agrupamientos y biagrupamientos, con el fin de obtener grupos de genes con patrones de expresión génica similar. Los grupos así obtenidos fueron validados estadística y semánticamente: la primera validación de acuerdo a las recomendaciones del algoritmo aplicado y la segunda utilizando la medida de similitud semántica de superposición de términos (term overlap) sobre la ontología génica GO, para garantizar que los grupos obtenidos tengan relevancia tanto estadística como biológica. Para cada grupo de genes válido se procedió a recuperar la región intergénica de sus integrantes, a las cuales se le aplicaron algoritmos de búsqueda de patrones de manera de determinar la existencia de posibles sitios regulatorios comunes a todo el grupo.

1. Introducción

1.1. La tuberculosis en humanos y animales de interés económico

Mycobacterium tuberculosis y *M. bovis* son los agentes causantes de tuberculosis en humanos y varias especies de ganado. Ambas especies tienen genomas muy similares y relativamente grandes (contando cada una con aproximadamente 4,000 “loci”), permitiendo un metabolismo sofisticado con la habilidad de responder a una variedad de estímulos fuera y dentro de la célula. En consecuencia, los experimentos con microarreglos son herramientas valiosas para estudiar la regulación de la transcripción del género Mycobacterium. Algunos de estos experimentos se pueden obtener desde el sitio del NCBI (National Center for Biotechnology Information) donde se encuentran depositados.

La existencia de microorganismos en nuestro planeta es ubicua y se los encuentra en los ambientes más extremos, desde aguas termales a temperaturas casi de ebullición hasta los polos. A lo largo de la evolución los microorganismos se adaptaron a distintos estilos de vida y localizaciones, por ejemplo, en los océanos, en el suelo o asociados a plantas o animales, tanto formando asociaciones mutuamente benéficas, conocidas como simbiosis, como estableciendo relaciones patogénicas. Las bacterias patógenas, debido a su impacto médico, social y económico, ocupan un lugar destacado en los programas de investigación, desarrollo de nuevos medicamentos y medidas de control sanitario. Entre las bacterias patógenas más conocidas se

encuentran aquellas que causan las diversas formas de tuberculosis que se manifiestan en humanos, bovinos porcinos y caprinos, entre otras especies animales. La tuberculosis humana nunca desapareció completamente, a pesar de que el agente causal, Mycobacterium tuberculosis, fue descubierto hace ya más de cien años, se cuenta con una vacuna que ofrece un buen grado de protección y durante mucho tiempo se contó con buenas terapias farmacológicas para controlar la infección. Aún peor, en los últimos años se registra un aumento en el número de casos totales y casos con desenlace fatal. Las causas de este aumento son varias, entre las que podemos mencionar: aparición de nuevas variantes de bacterias resistentes a antibióticos, desnutrición y el incremento de la población inmunosuprimida causada por la pandemia de SIDA.

La tuberculosis humana es responsable de 1,6 millones de muertes por año, lo que representa cerca del 11% de todas las muertes por infecciones. M. tuberculosis pertenece al grupo de las micobacterias, que incluye también a M. bovis, el agente causal de tuberculosis en bovinos, cabras, perros, gatos, búfalos, cerdos y algunas especies de simios entre otros. En el caso de los bovinos, la prevalencia (animales infectados sobre el total) en animales faenados en establecimientos controlados a nivel federal por SENASA descendió al 1,2% en 2005, desde un 6,7% en 1969 como consecuencia de un plan de control. El objetivo final es la erradicación y los avances son alentadores, pero se debe tener en cuenta que no todos los frigoríficos del país son controlados por SENASA y que animales con síntomas evidentes de la enfermedad no son enviados a faenar, pero pueden haber contagiado a

otros en los rodeos. La prevalencia de tuberculosis en cerdos, medida también como número de casos detectados en animales faenados en establecimientos bajo control del SENASA, es algo inferior al 1%. Con baja frecuencia M. bovis puede provocar enfermedad en humanos y una cepa atenuada de esta especie es la que se utiliza para producir la vacuna BCG.

Al igual que ocurre con cualquier otro ser vivo, la capacidad de las micobacterias para invadir un hospedador, evadir la respuesta de sus sistema inmune y desarrollar mecanismos para tolerar las terapias farmacológicas está codificado en su información genética. Por este motivo, desde hace muchos años la investigación genética en micobacterias patogénicas es un campo activo en medicina y veterinaria. Debido a la gran similitud genética entre M. bovis, el patógeno usual en bovinos, y M. tuberculosis, el responsable más frecuente de la tuberculosis humana, y a la mayor disponibilidad de herramientas de laboratorio y antecedentes experimentales para la segunda especie, es frecuente emplear cepas de M. tuberculosis como modelos para ambas enfermedades.

1.2. La información genética en la célula

El ácido desoxirribonucleico (ADN) contiene toda la información que un ser vivo necesita para su funcionamiento y desarrollo. Desde el punto de vista químico es una cadena larga formada por el ordenamiento secuencial de moléculas más sencillas conocidos como nucleótidos, formados por una base nitrogenada, un azúcar y un grupo fosfato. En el caso del ADN una

denominación más específica para sus nucleótidos es desoxirribonucleótidos, estos son de cuatro tipos, nombrados según la base nitrogenada: adenina (A), timina (T), citosina (C) y guanina (G). La molécula de ADN normalmente está formada por dos cadenas de ADN enfrentadas y organizadas en forma de doble hélice. En estas hélices los nucleótidos que pueden enfrentarse entre sí, o aparearse, son A con T y C con G. Esto determina que si uno conoce la secuencia de una cadena, fácilmente puede determinar la secuencia de la otra. Por ejemplo, la secuencia

A C C G T T A C C G

Va a estar apareada por una cadena con esta secuencia:

T G G C A A T G G C

Otro ácido nucleico importante es el ácido ribonucleico (ARN), a diferencia del ADN es una cadena simple, en lugar de doble y la molécula de azúcar que lo forma es la ribosa en lugar de la desoxirribosa. Los ribonucleótidos que lo forman son adenina (A), citosina (C), guanina (G) y uracilo (U), en lugar de timina.

Los genes son las unidades de información genética. En su gran mayoría cada uno de ellos codifica la información necesaria para producir una proteína. El conjunto de todos los genes que constituyen el material genético de un ser vivo se conoce como genoma. Cada proteína es una secuencia de aminoácidos de largo variable, desde aproximadamente 30 hasta alrededor de 800. Todas las proteínas se forman a partir de la combinación de 20 aminoácidos diferentes. Las proteínas son moléculas que intervienen en todos los procesos biológicos de una célula: algunas se ocupan de la

manipulación de la información genética, intervienen en todas las reacciones bioquímicas que lleva a cabo una célula, otras tienen funciones estructurales y unas cuantas intervienen en los procesos de detección y adaptación a los cambios ambientales.

Si bien los genes codifican la información para una proteína, la síntesis de éstas no ocurre directamente a partir de aquellos. Existe una molécula intermediaria, el ácido ribonucleico mensajero, mejor conocido por su sigla, ARNm. Existen ribosomas que copian el mensaje contenido en un gen a ARNm y luego éste se traduce a proteínas. Como se mencionó más arriba, las proteínas están compuestas por 20 aminoácidos diferentes, pero los nucleótidos son sólo cuatro. Para poder traducir un mensaje "escrito" en el lenguaje de los nucleótidos a aminoácidos, las células emplean "palabras" de tres letras, conocidas como codones.

La figura 1 muestra el código genético, que es la correspondencia entre codones y aminoácidos. Por ejemplo el codón UUU corresponde a fenilalanina (phe). Se puede observar que algunos aminoácidos están codificados por varios codones, por ejemplo la serina (ser), y que ciertos codones no codifican para ningún aminoácido, pero funcionan como señales de fin de mensaje (stop). El ejemplo siguiente muestra la correpondencia entre los últimos codones de una molécula de ARNm y una proteína:

UCU CCU GCU GCC UAG

ser tyr ala ala stop

		segunda base			
		U	**C**	**A**	**G**
primera base	**U**	UUU (Phe/F) UUC (Phe/F) UUA (Leu/L) UUG (Leu/L)	UCU (Ser/S) UCC (Ser/S) UCA (Ser/S) UCG (Ser/S)	UAU (Tyr/Y) UAC (Tyr/Y) UAA *Stop* UAG *Stop*	UGU (Cys/C) UGC (Cys/C) UGA *Stop* UGG (Trp/W)
	C	CUU (Leu/L) CUC (Leu/L) CUA (Leu/L) CUG (Leu/L)	CCU (Pro/P) CCC (Pro/P) CCA (Pro/P) CCG (Pro/P)	CAU (His/H) CAC (His/H) CAA (Gln/Q) CAG (Gln/Q)	CGU (Arg/R) CGC (Arg/R) CGA (Arg/R) CGG (Arg/R)
	A	AUU (Ile/I) AUC (Ile/I) AUA (Ile/I) AUG (Met/M)	ACU (Thr/T) ACC (Thr/T) ACA (Thr/T) ACG (Thr/T)	AAU (Asn/N) AAC (Asn/N) AAA (Lys/K) AAG (Lys/K)	AGU (Ser/S) AGC (Ser/S) AGA (Arg/R) AGG (Arg/R)
	G	GUU (Val/V) GUC (Val/V) GUA (Val/V) GUG (Val/V)	GCU (Ala/A) GCC (Ala/A) GCA (Ala/A) GCG (Ala/A)	GAU (Asp/D) GAC (Asp/D) GAA (Glu/E) GAG (Glu/E)	GGU (Gly/G) GGC (Gly/G) GGA (Gly/G) GGG (Gly/G)

Figura 1. El código genético. La figura muestra la correspondencia entre codones de tres letras de ribonucleótidos y aminoácidos.

El flujo típico y simplificado de información desde un gen a una proteína es:

gen (ADN) ⇨ RNAm ⇨ proteína

1.3. La regulación de la expresión génica

Los genomas de M. tuberculosis y M. bovis son muy similares entre sí y contienen aproximadamente 4.000 genes, que codifican todas las proteínas que necesitan estas bacterias para desarrollarse, dividirse, invadir sus hospedadores, adaptarse a los cambios del ambiente, etc. Estos 4.000 genes no se transcriben a ARNm al mismo tiempo, porque las proteínas para las que

codifican no se necesitan todas en todo momento. Por el contrario, muchas de ellas se necesitan en momentos muy específicos, por ejemplo, en un punto determinado del ciclo de división celular. Los seres vivos modulan la expresión de los genes para economizar recursos y mantener la coordinación entre procesos. Existen mecanismos de regulación que no solo determinan si un gen está apagado o no, es decir, si se va a expresar o no, sino que también controlan el nivel de expresión, esto es, la cantidad de mensajeros que se van a producir en un momento determinado. En muchos casos la regulación de la expresión de un gen se correlaciona con la cantidad de proteínas que se traducirán, lo que permite regular el proceso biológico en el que está involucrada esa proteína.

Existen proteínas, denominadas factores de transcripción, que regulan el nivel de expresión de los genes. Reconocen y se unen a secuencias cortas y específicas de nucleótidos en la región de ADN adyacente al inicio del gen, o promotor. Estas secuencias cortas son los llamados sitios de unión del factor de transcripción. En Mycobacterium se conocen numerosos factores de transcripción y la figura 2 muestra un esquema conceptual de su funcionamiento.

Conocer los factores de transcripción, sus sitios de unión y los genes que controlan es de suma importancia para descubrir las redes de interacción de la expresión génica. Esto puede servir, por ejemplo, para dilucidar el funcionamiento de mecanismos complejos de patogenia. Algunos sistemas de regulación de la transcripción están bien descriptos, pero quedan muchos por descubrir.

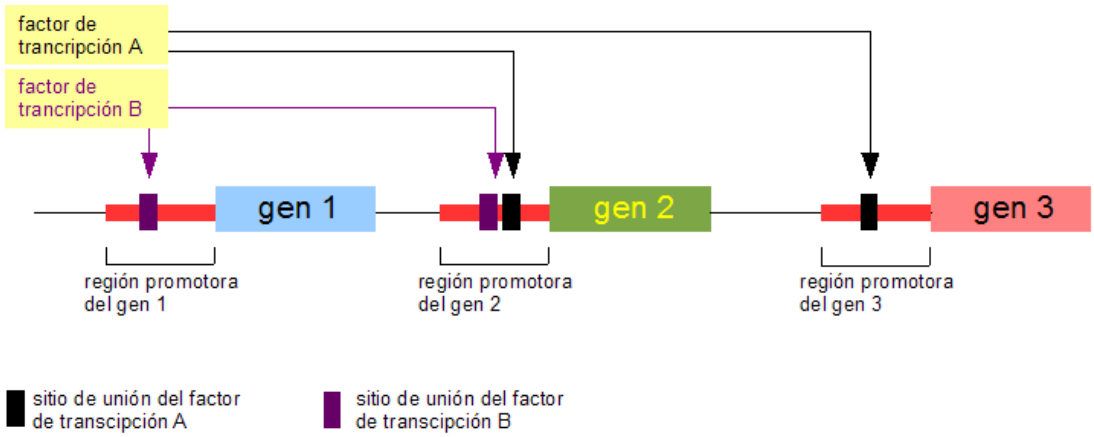


Figura 2. Esquema de la regulación de la expresión génica mediante factores de transcripción. El factor de transcripción A regula a los genes 2 y 3, y el factor B a los genes 1 y 2. El gen 2, en consecuencia, tiene un control doble de su transcripción

1.4. Análisis de la expresión génica con micromatrices de ADN

Una herramienta extremadamente útil para conocer el perfil de expresión global de un genoma es mediante el uso de microarreglos de ADN, que son soportes sólidos que contienen copias de todos o al menos muchos de los genes que conforman un genoma dispuestos en un ordenamiento matricial. Se toma una muestra de ARNm de la célula, se lo trata con una sustancia fluorescente y se lo enfrenta al microarreglo. Los mensajeros se unirán por complementariedad a moléculas similares. Luego se ilumina con una fuente de luz especial para revelar fluorescencia y se escanea la imagen. Los puntos del microarreglo que emiten fluorescencia se corresponden con genes que se están expresando. De esta forma, El uso de microarreglos permite conocer el nivel de expresión de cientos o miles de genes simultáneamente.

Una tecnología frecuentemente utilizada en el análisis de microarreglos es la de dos canales. Brevemente, se toma ARNm de un cultivo celular sometido a una condición experimental, por ejemplo, estrés por alta temperatura. El ARNm se "marca" con una sustancia fluorógena que emite en la longitud de onda del rojo. Simultáneamente se extrae ARNm de otro cultivo similar e independiente, que no fue sometido a la condición de estrés, es decir, la condición control. Esta segunda muestra de ARNm se marca con un fluorógeno que emite en la longitud de onda del color verde. Ambas muestras se enfrentan a un microarreglo. Los puntos de la matriz que contengan secuencias que corresponden a genes que se expresan en la condición de estrés y en la de control, emitirán luz amarilla. Aquellos que se expresan exclusivamente ante la situación de estrés, emitirán luz roja, y los que se expresan en la condición control, pero no ante el aumento de temperatura, emiten luz verde. El microarreglo se escanea y a partir del tratamiento digital de la imagen se obtiene una tabla indicando los niveles de expresión relativos.

En los experimentos reales se comparan al mismo tiempo varias condiciones experimentales y se incluyen varias réplicas por tratamiento, como se muestra en la figura 3. Teniendo en cuenta que con cada microarreglo se pueden interrogar varios miles de genes, es evidente que los análisis de expresión génica empleando microarreglos generan una elevada cantidad de datos, adecuados para diferentes tipos de análisis. Los más comunes son la búsqueda de genes con expresión diferencial y el agrupamiento de genes con perfiles de expresión similares. El análisis de datos de microarreglos se

convirtió en un campo muy fecundo para la aplicación de técnicas de minería de datos y descubrimiento del conocimiento.

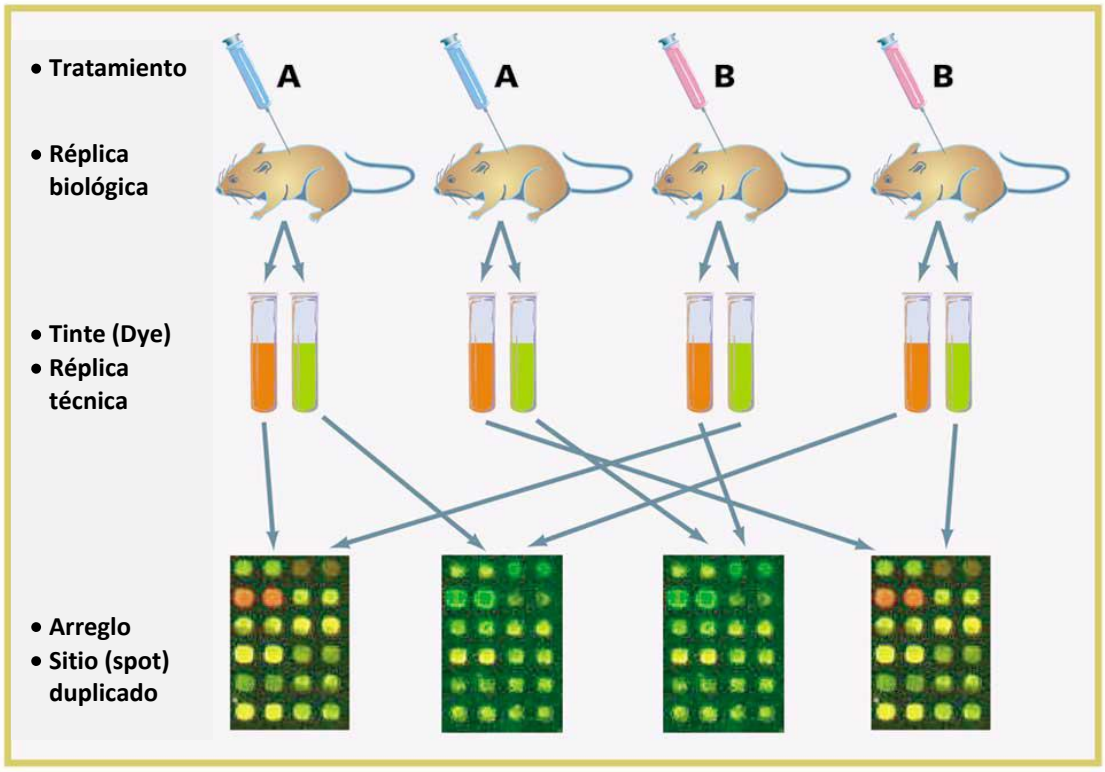


***Figura 3 Una representación esquemática de las tres capas de diseño de un experimento con micro arreglo simple. En la capa superior, las unidades biológicas (ratones) son asignadas* a grupos de tratamientos diferentes (A y B). En la capa del medio, dos muestras de ARN son obtenidas de cada uno de los ratones. Estas réplicas son marcadas diferencialmente e "hibridizadas" en pares a micro arreglos. Cada par involucra una comparación directa de un ratón A y de un ratón B, y las marcas con tintes (dye labels) son invertidas en dos de las cuatro comparaciones. La capa inferior del experimento es representada por las imágenes de los arreglos, en las cuales los tientes duplicados de clones son evidentes.**

1.5. Repositorio de datos

El NCBI (National Center for Biotechnology Information) provee acceso a diferentes repositorios de información biomédica y genómica. Uno de estos repositorios es el GEO (Gene Expression Omnibus). GEO almacena datos

curados (validados por el NCBI) de datos de expresión génica usando el estándar MIAME.

A este repositorio se accede desde el sitio http://www.ncbi.nlm.nih.gov/geo/ donde los usuarios pueden especificar las consultas a fin de obtener datos e información referidos a los experimentos requeridos.

El estándar MIAME (*Minimum Information About a Microarray Experiment*) especifica la mínima información que debería ser incluida cuando se describe un experimento con microarreglos. Lo seis elementos más críticos incluidos en MIAME son:

1. Los datos en bruto para cada *hibridización.*
2. El dato final procesado (normalizado) para el conjunto de hibridizaciones en el experimento.
3. Las anotaciones esenciales, incluyendo factores experimentales y sus valores.
4. El diseño del experimento, incluyendo relaciones entre los datos de la muestra.
5. Anotaciones del microarreglo.
6. Los protocolos de laboratorio y procesamiento de datos esenciales

Dentro del sitio GEO las consultas pueden ser especificadas de diversas maneras: texto libre, palabras claves y especies, tipo de estudio, autor, fecha o plataforma entre otras. En nuestro caso, se realizaron búsquedas de conjunto de datos para el organismo *Mycobacterium*, recuperando de esta manera varios de los experimentos disponibles para dicho organismo. Una

vez seleccionados los experimentos con los cuales se desea trabajar se puede realizar la consulta utilizando el código que identifica al experimento. El resultado de la búsqueda y su estructura se pueden apreciar en el gráfico de la figura 4, en donde se puede distinguir que cada experimento consta de una serie de muestras, cada una con sus tablas de resultados y su plataforma asociada (puede darse el caso que exista más de una plataforma[1] asociada a las muestras, en cuyo caso se debería tener cuidado de cómo comparar resultados correspondientes a las distintas plataformas). Primero se muestra el nombre del experimento, con la descripción del estudio; posteriormente las muestras, en donde se describen las condiciones bajo las cuales un individuo fue tratado; y finalmente la plataforma utilizada.

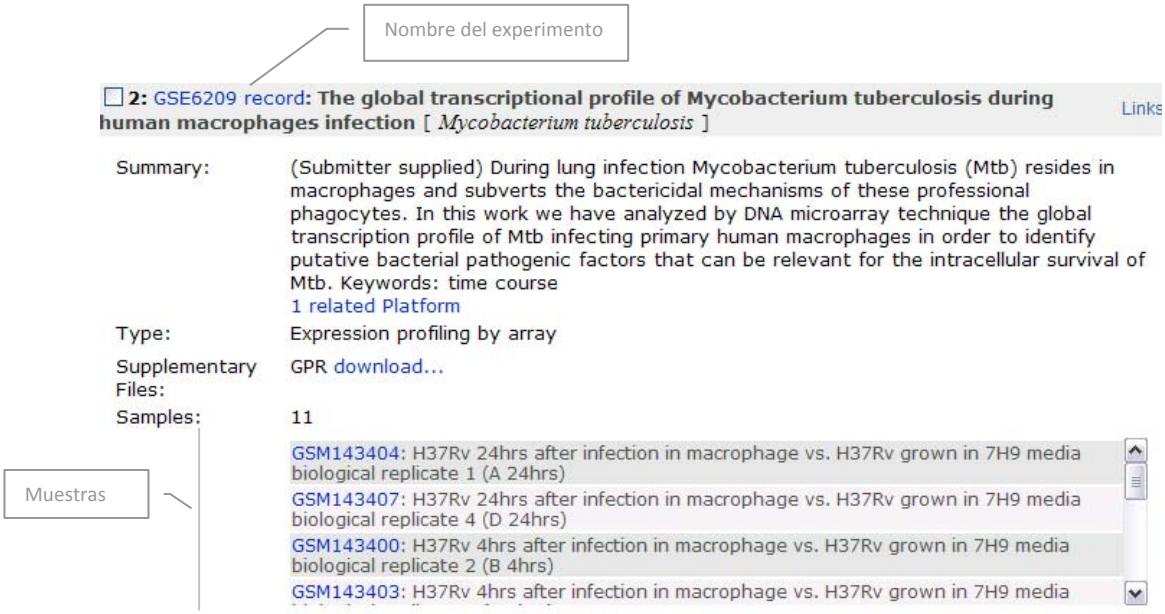


Figura 4 (a). Resultado de la búsqueda del experimento GSE6209 en el sitio GEO. Se puede apreciar que el experimento está compuesto de 11 muestras.

[1] Una plataforma es la combinación de marca y modelo de escáner, diseño del microarreglo y tipo de reacciones involucradas en la reacción de detección.

Sample GSM143404 Query DataSets for GSM143404

Status	Public on Mar 27, 2008
Title	H37Rv 24hrs after infection in macrophage vs. H37Rv grown in 7H9 media biological replicate 1 (A 24hrs)
Sample type	RNA

Channel 1

Source Name	H37Rv grown in 7H9 media
Organism	Mycobacterium tuberculosis
Characteristics	strain H37Rv Time: Control exponentially growing H37Rv in 7H9 media
Treatment protocol	NONE

Channel 2

Source Name	H37Rv 24hrs after infection in macrophage (THP1)
Organism	Mycobacterium tuberculosis
Characteristics	strain H37Rv Time: 24hrs after infection

Figura 4 (b). Una de las muestras con sus dos canales.

Platform ID	GPL4057	
Series (2)	GSE6209	The global transcriptional profile of Mycobacterium tuberculosis during human macrophages infection
	GSE7963	Mycobacterium tuberculosis and macrophage response

Data table header descriptions

ID_REF	
VALUE	Print Tip Lowess Normalized Log 2 ratio test/reference
Flags	0 if good, -50 not found
Dia.	spot diameter
F635 Median	Cy5 Intensity
B635	Cy5 Background intensity
F532 Median	Cy3 Intensity
B532	Cy3 Background intensity

Data table

ID_REF	VALUE	Flags	Dia.	F635 Median	B635	F532 Median	B532
010101	-1.91896	0	110	221	53	837	100
010102	-2.51291	0	100	165	56	943	126
010103	-2.36441	0	110	278	60	1431	147
010104	-1.28526	0	100	169	64	412	169

Platform GPL4057 Query DataSets for GPL4057 ← Plataforma

Status	Public on Aug 04, 2006
Title	PHRI-UMNDJ Mycobacterium tuberculosis 4.8K CAG_Mtb
Technology type	spotted oligonucleotide
Distribution	non-commercial
Organism	Mycobacterium tuberculosis
Manufacturer	Center for Applied Genomics
Manufacture protocol	Oligonucleotide 70mer long from the OPERON TB set version 1.0 are spotted using an omnigrid arrayer at a concentration of 25uM in 3x SSC onto glass slides coated with Poly-L-Lysine.
Support	glass
Coating	polysine

Figura 4 (c). Plataforma del experimento.

Data table header descriptions

ID	
Rv-ORF_ID	Rv or ORF number
Name	Gene name
ORF	
SPOT_ID	

Data table

ID	Rv-ORF_ID	Name	ORF	SPOT_ID
010101	Rv0098	Rv0098	Rv0098	
010102	Rv0100	Rv0100	Rv0100	
010103	Rv0102	Rv0102	Rv0102	
010104	Rv0104	Rv0104	Rv0104	
010105	Rv0106	Rv0106	Rv0106	
010106	Rv0108c	Rv0108c	Rv0108c	
010107	Rv0122	Rv0122	Rv0122	

Figura 4 (d). Los resultados de las muestras poseen referencias a los genes pertenecientes a la plataforma a través de un identificador o ID. Hay que recurrir al ID para saber a qué gen corresponde cada medición. En la figura se muestran las primeras 7 filas de la tabla que relacionan el gen con el ID.

Los experimentos seleccionados y recuperados desde GEO para el presente trabajo son los siguientes:

Tabla de experimentos seleccionados

Identificador	Descripción	# muestras
GDS2677	**Efecto de la capreomycin sobre *Mycobacterium tuberculosis***	4
GSE6209	***Mycobacterium tuberculosis* y la respuesta de los macrófagos**	11
GSE8639	**Regulación de *Mycobacterium tuberculosis* hipóxico gen que codifica la respuesta alfa-cristalina**	6
GSE10391	**In vitro dormancia alcanzado por múltiples tensiones en *Mycobacterium tuberculosis***	75
GSE12364	**Respuesta de transcripción global a la vancomycin en *Mycobacterium tuberculosis***	12
GSE15976	**Respuesta Sigma factor B (sigB) condiciones de estrés (0.05% SDS and 5mM Diamide)**	36
GSE365	**Reparación del ADN en Mycobacterium tuberculosis**	28
GSE7962	**Sigma factor de E de *Mycobacterium tuberculosis* controla la expresión de los componentes**	23

	bacterianos que modulan macrófagos	
GSE9776	**Efecto del tratamiento con INH en la expresión de genes de *Mycobacterium tuberculosis* en varios modelos de inactividad**	17

Tabla 1. Experimentos de microarreglos seleccionados, donde se observa el identificador del experimento, la descripción y el número de muestras.

Dentro de cada experimento el repositorio permite acceder a un detalle del mismo, en donde se presentan cada una de las muestras que lo componen. Por ejemplo, el experimento GDS2677 tiene cuatro muestras, donde se puede apreciar que hay dos condiciones, cada una de las cuales se replican dos veces (las réplicas se identifican por lo general mediante el mismo nombre seguido de un post-fijo que indica el número de la réplica, en este caso rep. 1 y rep. 2, para cada una de las condiciones). Además, en el mismo sitio se puede obtener un agrupamiento básico, junto a una representación en forma de mapa de calor, como se aprecia en la figura 5.

Muestras del experimento GDS2677	
Muestra	Descripción
GSM183531	**M. tuberculosis_H37Rv_log-phase_rep1**
GSM183633	**M. tuberculosis_H37Rv_log-phase_rep2**
GSM183632	**M. tuberculosis_capreomycin_4h_rep1**
GSM183634	**M. tuberculosis_capreomycin_4h_rep2**

Tabla 2. Muestras para el experimento GDS2677, donde se observa el identificador de la muestra y su descripción.

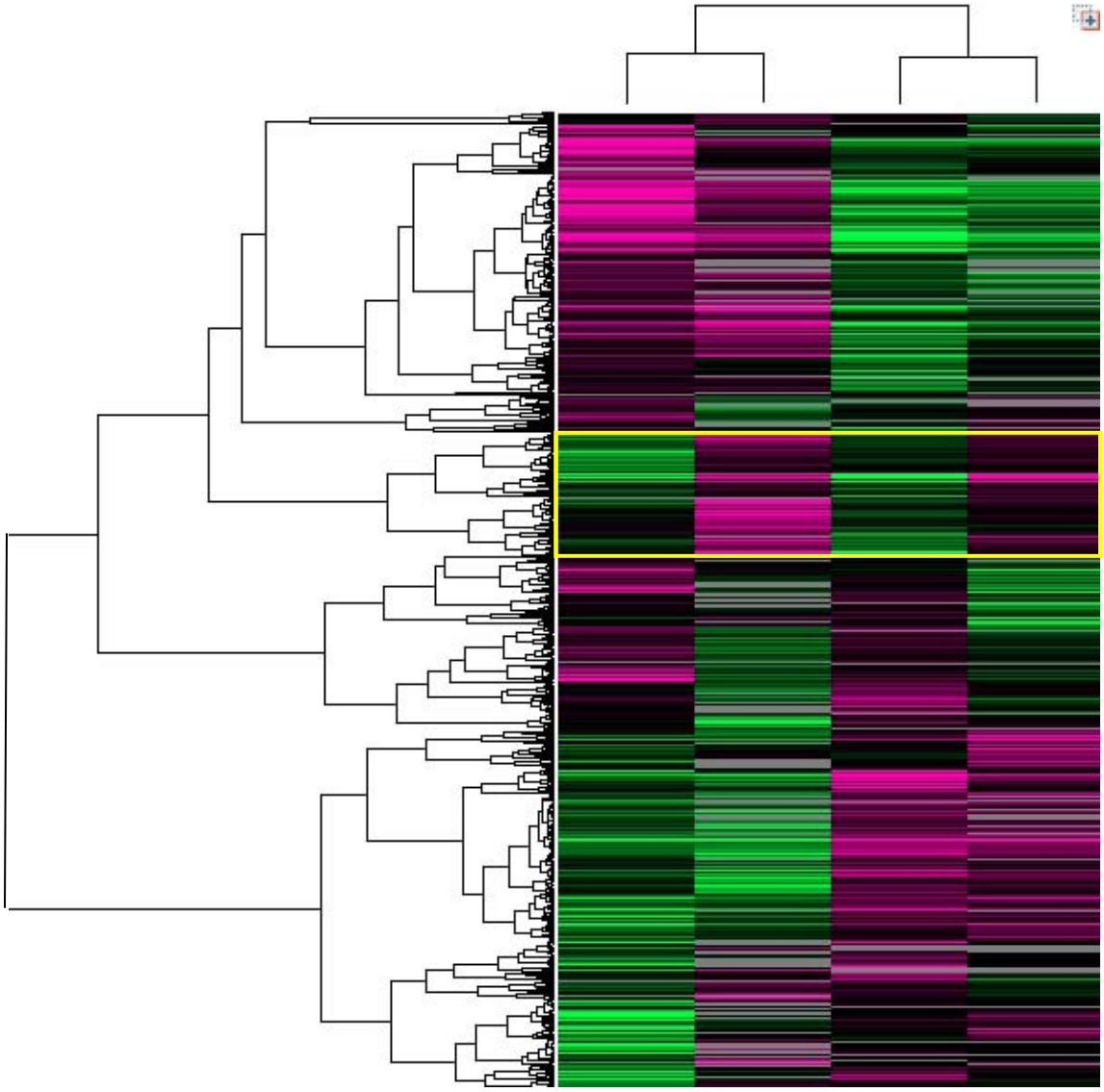


Figura 5. Agrupamiento del experimento GDS2677. Puede apreciarse la complejidad de los datos de estos experimentos si observamos que para las repeticiones bajo las mismas condiciones el mapa de calor[2] correspondiente no siempre guarda correlato. Por ejemplo, ver en la sección señalada con el rectángulo amarillo: para la misma condición se sobre expresa una de las réplicas, mientras se sub expresa la otra réplica. Los mismo pasa para las réplicas bajo la otra condición

La información recuperada para cada experimento se almacena, para su procesamiento, en una estructura de datos matricial, donde las filas son los

[2] Un mapa de calor es una representación gráfica de los datos en donde los valores contenidos en una matriz son representados mediante colores.

genes y las columnas los tratamientos. En la figura 6(a) se presenta una muestra de los datos de la matriz para el experimento GSE6209, con 6 genes y 4 tratamientos (de un total de 3924 genes y 11 tratamientos). Debajo, en la figura 6(b), un diagrama de caja para el conjunto total de datos del mismo experimento. Se almacena cada experimento en una matriz distinta, y se analizan cada una de estas matrices de manera independiente.

	GSM143399	**GSM143400**	**GSM143401**	**GSM143402**
RV0001	0,967118	1,82321	1,11132	2,0513
RV0002	0,38549	1,46465	0,505222	-0,0613247
RV0003	0,189768	0,464284	-0,375436	-0,980485
RV0004	-0,420102	-0,253198	-0,164523	-0,672465
RV0005	0,350077	0,21535	-0,179201	0,271957
RV0006	-0,400041	-0,281939	0,507833	-0,234

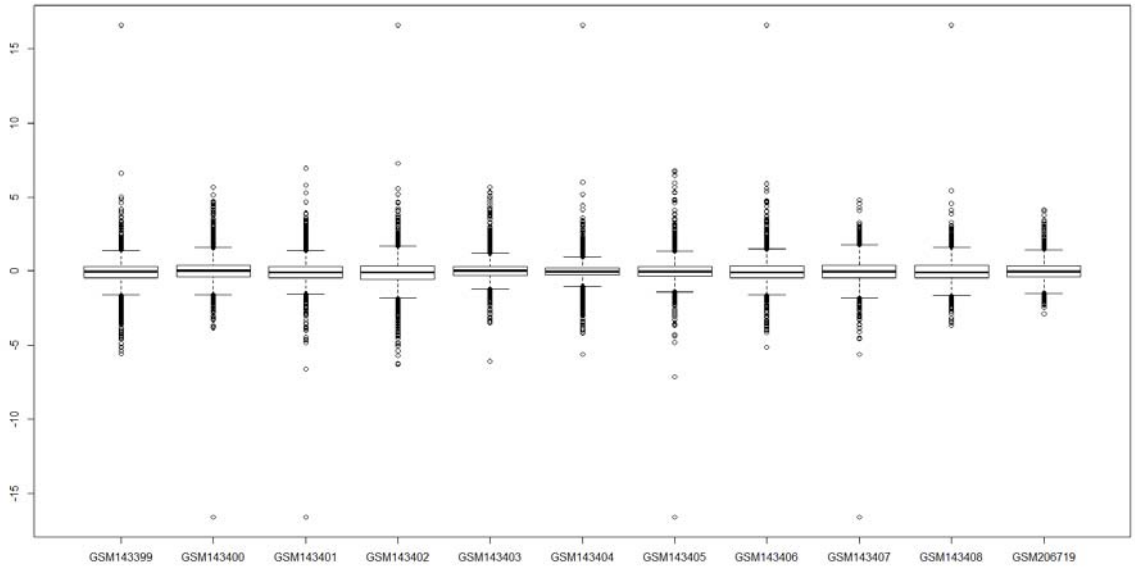


Figura 6(a). Muestra de la matriz para el experimento GSE6209, donde se visualizan los valores de 6 genes, de un total de 3924, para 4 de los 11 tratamientos. Figura 6(b). Diagrama de caja para los datos completos del experimento GSE6209.

Otro repositorio de datos utilizado en este trabajo es la base de datos DOOR (Database for prOkaryotic OpeRons), que contiene operones (obtenidos mediante predicciones computacionales) de todos los genomas procariotas secuenciados. Un operón es un conjunto de genes adyacentes que se transcriben en una molécula de mRNA única y su regulación es coordinada. Presentan por tanto un promotor único. Frecuentemente los genes que integran un operon participan del mismo proceso biológico.

El promotor de un gen es la región de ADN que controla la iniciación de la transcripción de dicho gen a ARN. Dicha región está compuesta por una secuencia específica de ADN localizado justo donde se encuentra el punto de inicio de la transcripción del ADN y contiene la información necesaria para activar o desactivar el gen que regula. En las regiones promotoras es común encontrar patrones cortos de secuencias de nucleótidos que se repiten en diferentes promotores y se denominan motivos (motifs en inglés).

Los motivos también pueden ser pensados como secuencias generalizadas y representados como matrices de puntajes. Mientras una secuencia tiene una letra a cada posición, un motivo tiene un vector de valores con un puntaje para cada letra del alfabeto (4 valores en el caso del ADN). Luego el puntaje de una subsecuencia, dentro de una secuencia dada, con respecto a su emparejamiento con el motivo, estará dado por la suma de los puntajes, de acuerdo al valor otorgado a cada una de sus coincidencias, como puede observarse en la figura 7.

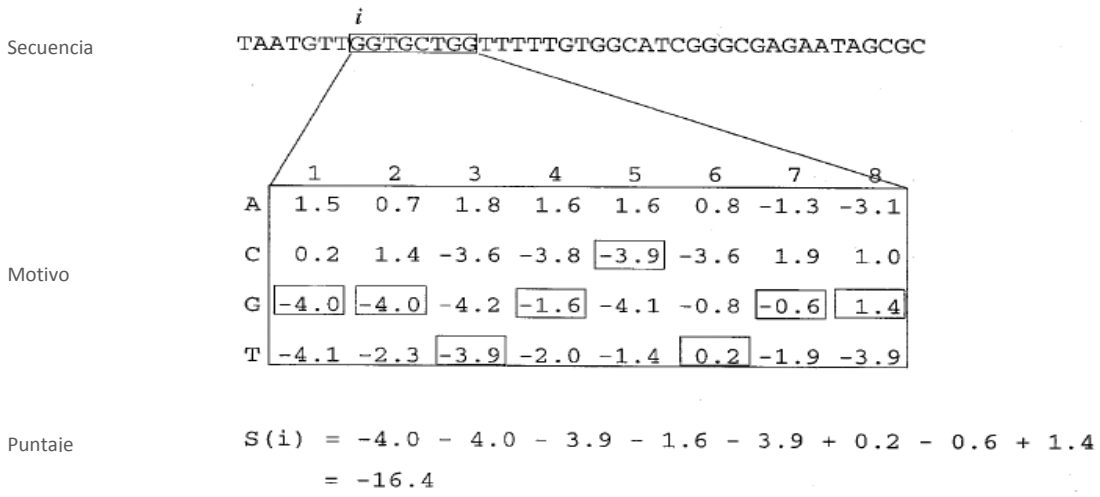


Figura 7. Matriz de puntajes representando un motivo. Para obtener el puntaje se suma cada elemento de la secuencia en el motivo.

1.6. Normalización de los datos

La normalización es un término general para una colección de métodos que se dirigen a resolver los errores sistemáticos y el sesgo introducidos por los experimentos con microarreglos. En este paso también se realizan las transformaciones necesarias para llevar la distribución de los datos a una distribución esperada. Posteriormente a la limpieza de los datos, en donde se remueven las características[3] de mala calidad y se substrae el fondo, entre otras cosas, se procede a realizar los pasos de normalización dentro de los arreglos (*within array normalization*) y entre los arreglos (*between array normalization*).

[3] Los algoritmos de extracción de características convierten la imagen del microarray en información numérica que cuantifica la expresión de los genes. El procesamiento de imágenes involucrado en esta extracción tiene un impacto importante sobre la calidad del dato y la interpretación que se puede hacer sobre este.

Como se mencionó anteriormente, en los experimentos de dos canales, cada canal se marca con dos colorantes fluorescentes diferentes en dos reacciones químicas separadas, y su intensidad es medida con dos diferentes laser operando a dos longitudes de onda diferentes. Además, el material que constituye el soporte del microarreglo puede presentar heterogeneidad sobre su superficie. Cuando se mide la expresión diferencial entre las dos muestras, es necesario asegurar que las mediciones representen verdaderamente una expresión diferencial del gen, y no una tendencia o error introducido por el método experimental. Se debe ser capaz de comparar las intensidades de Cy3 y Cy5 eliminando cuatro fuentes de sesgos sistemáticos:

1. Las etiquetas (*labels*) para Cy3 y el Cy5 pueden estar diferencialmente incorporados dentro del ADN en abundancia diferente.
2. Los colorantes Cy3 y Cy5 pueden tener diferentes respuestas de emisión a la excitación del laser a diferentes abundancias.
3. Las emisiones de Cy3 y Cy5 pueden ser diferencialmente medidas.
4. Las intensidades medidas de los Cy3 y Cy5 sobre diferentes áreas del arreglo pueden ser diferentes debido a las heterogeneidades del arreglo.

Se pueden aplicar una serie de métodos con el fin de resolver estas consideraciones, donde los más utilizados son los dos que se detallan a continuación:

1. Gráfico de dispersión entre Cy5 y Cy3

El método más simple para constatar cuando los canales Cy3 y Cy5 se están comportando de una manera comparable es vía el gráfico de dispersión de los dos canales. Si esto sucede, la nube de puntos sobre el gráfico debería aproximarse a una línea recta. Esto es debido a que se espera que la mayoría de los genes se comporten de manera similar en ambos canales, y solamente se vean diferencialmente expresados un porcentaje menor.

2. Gráfico de dispersión del log ratio versus la intensidad promedio

   Otra forma de visualizar lo anterior es producir un gráfico de dispersión del logaritmo del ratio entre la intensidad de cada canal (log ratio), contra la intensidad promedio, para cada característica. Estos gráficos son referenciados como MA en la literatura de micro arreglos, y en ellos cada punto representa una característica, con la coordenada X tomando el valor promedio del log de las intensidades de Cy5 (canal rojo) y Cy3 (canal verde), y la coordenada Y la diferencia entre el log de intensidades de los canales de Cy5 y Cy3. El MA deriva su nombre debido a que la intensidad promedio es referida como A y el log ratio como M.

$$M = log_2\left(\frac{I}{G}\right) = log_2(R) - log_2(G)$$

$$A = \frac{1}{2} log_2(R\ G) = \ 1/2(log_2(R) + log_2(G))$$

Si los dos canales están respondiendo similarmente, luego el dato debería aparecer simétricamente al rededor de una línea horizontal que pase por el origen de coordenadas. En la figura 8 se muestra un ejemplo de un gráfico MA.

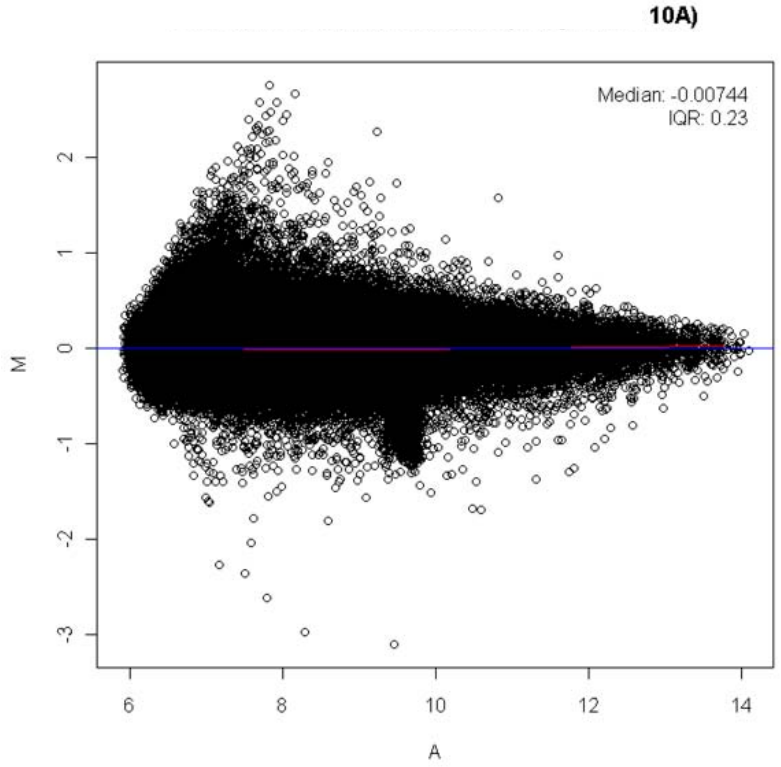


Figura 8. Ejemplo de gráfico MA

1.7. Agrupamientos

Un agrupamiento es el proceso de organizar objetos en grupos, o más precisamente, una partición de un conjunto de datos en subconjuntos, de tal manera que los datos en cada subconjunto compartan algún rasgo, probablemente proximidad de acuerdo a alguna medida de distancia definida. Los métodos de agrupamiento son técnicas habituales para el análisis estadístico de datos, los cuales son usados en muchos campos,

incluyendo minería de datos, reconocimiento de patrones, análisis de imágenes y bioinformática.

Existen tres pasos involucrados en el análisis de agrupamientos. El primero es el de preprocesamiento, en el cual se realizan un número de transformaciones de datos incluyendo selección de atributos, normalización y la elección de una función de distancia, para asegurar que los datos relacionados se agrupen. El segundo paso consiste en la selección y ejecución de uno o varios métodos de agrupamientos. En este paso es necesario además seleccionar los parámetros adecuados, dependiendo del método elegido. En el tercer paso se evalúan las particiones resultantes mediante técnicas de validación (figura 9).

Los métodos de agrupamientos se pueden dividir básicamente en **jerárquicos** y **no jerárquicos**. Los métodos jerárquicos encuentran sucesivos agrupamientos utilizando agrupamientos previamente establecidos. Estos pueden ser aglomerativos, comenzando con cada elemento como un agrupamiento separado y uniéndolos sucesivamente en agrupamientos mayores; o divisivos, comenzando con el conjunto de datos completo y dividiéndolos sucesivamente en agrupamientos más pequeños.

Los métodos no jerárquicos comienzan tanto desde una partición inicial de elementos o desde un conjunto inicial de “semillas” (elegidas con algún criterio particular o al azar entre los elementos del conjunto de datos) las cuales forman el centro de cada agrupamiento. El método no jerárquico más popular es el k-media, el cual comienza con una partición de k agrupaciones iniciales, asigna cada elemento del conjunto de datos a la agrupación cuyo centro sea el más cercano, y repite esta última operación hasta que no haya más reasignaciones.

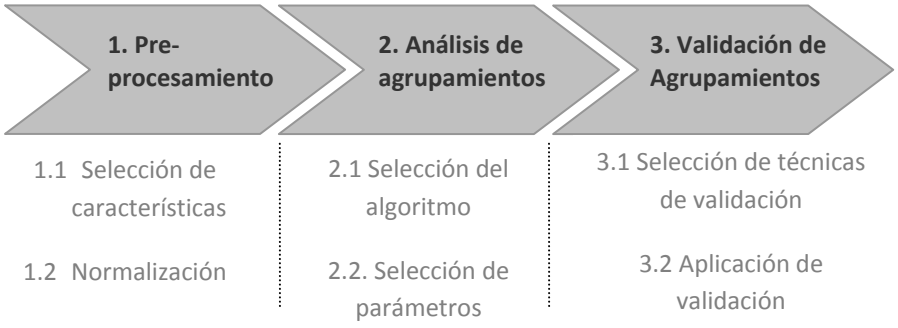


Figura 9. Pasos en el análisis de agrupamiento

A continuación se mencionan algunos métodos de agrupamiento utilizados en el presente trabajo.

1.7.1. K-means

El método K-means es uno de los algoritmos no supervisados más simples que resuelve el problema de agrupamientos. Sigue un procedimiento donde clasifica un conjunto de datos de manera sencilla en un cierto número k de grupos, valor que se toma como parámetro del algoritmo.

La idea es definir k centroides, uno para cada grupo. Inicialmente estos centroides se deben ubicar de alguna manera astuta, debido a que cada distribución puede causar un resultado diferente. La mejor elección es ubicarlos tan alejados unos de otros como sea posible. El siguiente paso es tomar cada elemento del conjunto de datos inicial y asignarlo al centroide más cercanos (de acuerdo a alguna medida de distancia predefinida). Cuando todos los puntos han sido asignados, el primer paso se ha completado con un primer agrupamiento. Seguidamente se procede a calcular los k centroides

(uno para cada grupo) y nuevamente se asignan todos los elementos a su centroide más cercano. Este procedimiento se repite de manera de minimizar la siguiente función objetivo:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2$$

donde $||x_i^{(j)} - c_j||$ es la medida de distancia elegida entre el elemento $x_i^{(j)}$ y el centroide c_j.

Aunque se puede demostrar que el procedimiento siempre termina, el algoritmo no siempre encuentra necesariamente la configuración óptima. El algoritmo es también sensible a los centros inicialmente elegidos. Para reducir este efecto el K-means puede ser ejecutado varias veces.

1.7.2. PAM

PAM es el acrónimo en inglés de Partition Around Medoids **[REY1992] y [KAU1990].**

El algoritmo busca k objetos representativos (medoides) que se encuentran centrados en los conglomerados que ellos definen. El medoide, objeto representativo del conglomerado, es aquel objeto para el cual la disimilitud promedio con todos los objetos en el conglomerado es mínima. El algoritmo PAM minimiza la suma de disimilitudes en vez de la disimilitud promedio.

La selección de k medoides se lleva a cabo en dos fases. En la primera, se obtiene un conglomerado inicial con la selección sucesiva de objetos representativos hasta hallar k objetos. El primer objeto es aquel para el cual la suma de las disimilitudes con todos los otros objetos es tan pequeña como

sea posible. En cada paso, PAM selecciona el objeto que hace decrecer la función objetivo (suma de disimilitudes) tanto como sea posible. En la segunda fase, se hace un intento de mejorar el conjunto de objetos representativos. Esto se hace al considerar todos los pares de objetos i, h para los cuales se ha elegido el objeto i y no el objeto h, realizando el reemplazo si se verifica que la elección de h en lugar de i hace decrecer la función objetivo. Este paso se repite hasta que la función objetivo alcance su mínimo. El hecho de minimizar la suma de las disimilitudes en lugar de la suma de los cuadrados de las distancias, es lo que hace a este método más robusto. Además tiene una forma de visualizar el resultado, la cual permite seleccionar el número de grupos óptimos.

1.7.3. CLARA

CLARA es el acrónimo en inglés de Clustering Large Applications **[KAU1990]** y puede tratar con conjuntos de datos mucho más grandes que otros métodos. Para conseguir esto el método considera subconjuntos de tamaño fijo de tal forma que el tiempo y el almacenamiento requerido es lineal con respecto a la cantidad de elementos del subconjunto.

Internamente, CLARA tiene dos pasos. Primero se toma una muestra del conjunto de objetos, y se divide en k conglomerados con el mismo algoritmo de PAM. A continuación, cada objeto que no pertenezca a la muestra se asigna al más cercano entre los k objetos representativos. La calidad de este conglomerado se define como la distancia promedio entre cada objeto y su objeto representativo. Este procedimiento se repite varias veces, y finalmente el agrupamiento con la distancia promedio entre cada objeto y su medoid es seleccionado como el agrupamiento definitivo.

De esta manera CLARA conserva las virtudes del método PAM con el agregado de poder trabajar con conjuntos de datos más grandes.

1.7.4. HOPACH

HOPACH es el acrónimo en inglés de Hierarchical Ordered Partitioning and Collapsing Hybrid **[KAU1990]**. Este método construye una jerarquía de agrupamientos. Para conseguir esto realiza particiones del conjunto inicial en forma recursiva utilizando el algoritmo PAM, a la vez que ordena y posiblemente une grupos en cada nivel. El algoritmo utiliza el criterio Mean/Median Split Silhouette o MSS para identificar los niveles de la jerarquía con los agrupamientos de máxima homogeneidad.

HOPACH es un método jerárquico que es un híbrido entre un algoritmo aglomerativo (bottom up) y un algoritmo divisivo (top down). El árbol del HOPACH se construye desde la raíz hacia las hojas, sin embargo a cada nivel los grupos similares pueden ser colapsados. Además, los grupos a cada nivel se ordenan con el mismo algoritmo determinístico sobre la misma métrica de distancia que es usada en el agrupamiento. De esta manera, el ordenamiento producido en el nivel final del árbol no depende sobre el orden del conjunto de datos inicial, como puede ser el caso con algoritmos que tienen un componente aleatorio en sus métodos de ordenamiento. A diferencia de otros métodos jerárquicos, los nodos del árbol de grupos no necesitan ser binarios, pudiendo existir más que dos descendientes en cada partición. El paso de división del algoritmo se realiza utilizando PAM y el criterio de *Median Split Silhoutte* (MSS) se utiliza para determinar el número óptimo de descendientes en cada nodo, para decidir cuáles pares de grupos colapsar a cada nivel y para identificar el primer nivel árbol con grupos homogéneos maximales. En cada caso el objetivo es minimizar el MSS, el cual es una medida de heterogeneidad del agrupamiento.

1.8. Consenso de agrupamientos

El gran reto en la combinación de resultados de agrupamientos es la definición de una función de consenso apropiada, capaz de mejorar los resultados de los agrupamientos individuales **[PON2011].**

El problema de la combinación de resultados de agrupamientos puede ser formalizado de la siguiente manera:

Siendo O = {O1,O2, … ,On} el conjunto de objetos, una combinación de agrupamientos es un conjunto P = {P1, P2,… , Pm }, donde Pi es una partición de los objetos O, para todo i =1, 2, … ,m donde el objetivo principal es encontrar una nueva partición P* de los datos a partir de las particiones en P. Para encontrar esta partición es necesaria una función de consenso, la cual es la encargada de combinar toda la información existente en el conjunto de las particiones P en una partición final P*.

Una buena estrategia de combinación, debe permitir encontrar nuevas estructuraciones más consistentes que las existentes. Esta estructuración de consenso debe ser además, lo más invariante posible a pequeñas variaciones en los datos, es decir, debe ser suficientemente robusta ante información ruidosa. Éstas, entre otras propiedades, son planteadas por los autores de algoritmos de combinación de resultados de agrupamientos **[PON2011]**; sin embargo no existe un criterio común ni una formalización rigurosa de las características que debe tener el consenso, más bien, cada autor propone las propiedades que cree que debe cumplir un buen mecanismo de combinación.

1.9. Biagrupamiento (Bicluster)

Dado un conjunto de expresiones de genes, organizado como una matriz con filas correspondientes a genes y columnas a condiciones, un análisis común es agrupar condiciones y genes en subconjuntos que tengan algún significado biológico. Esta tarea se traduce al problema computacional conocido como agrupamiento.

El análisis de agrupamiento hace varias asunciones que pueden no ser adecuadas en todas las circunstancias. Primero, el agrupamiento puede ser aplicado o bien a genes a bien a condiciones, y si, por ejemplo, se realiza sobre los genes, el grupo al cual pertenece un gen dado debe cubrir a todas las condiciones sin excepción (esto particularmente para el caso del trabajo presente, en donde existen diferentes tratamientos, puede ser que un grupo de genes responda de manera similar para un grupo de tratamientos y no par otros). Segundo, los algoritmos de agrupamiento usualmente buscan una cobertura disjunta de los elementos, requiriendo que ningún gene o condición vaya a más de un grupo.

La noción de biagrupamiento ofrece más flexibilidad a estos dos aspectos, como se muestra en los grupos de la figura 10. Los agrupamientos se realizan a la vez tanto por genes como por condiciones, donde no necesariamente si un gen pertenece a un grupo todas las condiciones para ese gen deben pertenecer también al grupo. También se relaja la condición donde un gen pertenece solamente a un grupo, como puede apreciarse en la superposición de dos bi-agrupamientos. Al no existir ninguna restricción apriori sobre la organización del biagrupamiento, genes y condiciones pueden ser parte de más de de un grupo o bien no pertenecer a ningún grupo.

Esta falta de restricciones permite gran libertad, pero es consecuentemente más vulnerable a sobreajuste, y se debe garantizar que la solución hallada sea significativa. En nuestro caso, la solución es significativa si los genes pertenecientes al biagrupamiento se expresan de manera similar para todos los tratamientos pertenecientes al biagrupamiento, de acuerdo a medidas que se explicarán en breve. Debido a este fenómeno es que la solución es acompañada de métodos de score basados en modelos estadísticos o heurísticos que definen cuál de las posibles submatrices tienen un significado real.

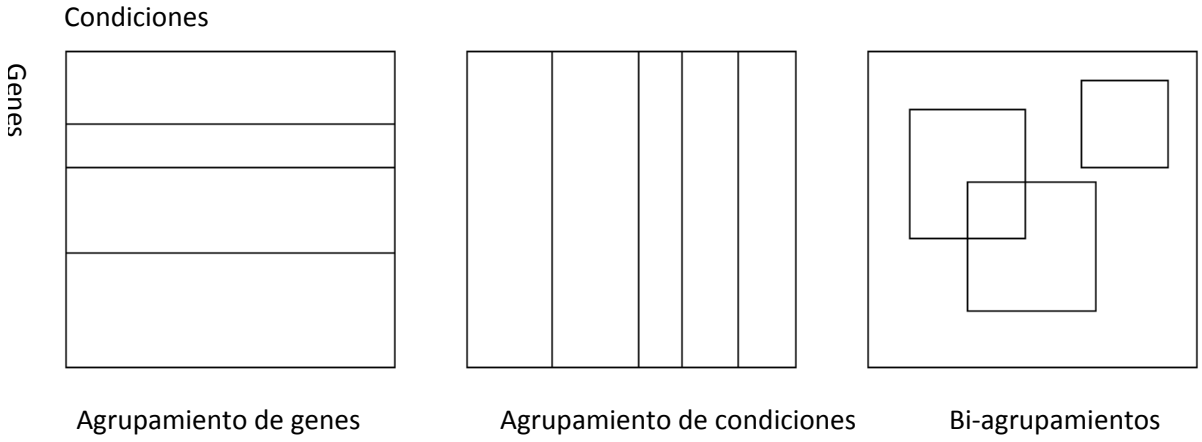


Figura 10. Agrupamiento y bi-agrupamiento. Agrupamientos corresponden a cortes disjuntos en la matriz. El gene en un grupo dentro del agrupamiento por genes debe con tener todas las columnas, y la condición en un grupo dentro del agrupamiento por condiciones debe contener a todos los genes. En cambio bi-agrupamientos corresponden a subconjuntos arbitrarios de filas y columnas.

Un criterio interesante para evaluar algoritmos de biagrupamiento consiste en identificar el tipo de biagrupamiento que el algoritmo es capaz de descubrir. En [MAD2004] identifican cuatro tipos de biagrupamientos:

1. Biagrupamientos con valores constantes.

   Los algoritmos más simples de biagrupamiento identifican subconjuntos de filas y de columnas con valores constantes, como es el caso de la figura 11(a). Los valores del biagrupamiento respetan la siguiente ecuación.

   $$m_{i,j=C}$$

   donde mij es el biagrupamiento con subíndices i ≤ n, j ≤ m y c es una constante.

2. Biagrupamientos con valores constantes sobre filas y columnas

   Este enfoque identifica subconjuntos tanto de filas o de columnas constantes, como son los casos de las figuras 11(b), en donde se

observan filas constantes, y 11(c), en donde se observan columnas constantes. Para este caso de valores constantes sobre filas los valores del biagrupamiento respetan alguna de las siguientes ecuaciones:

$$m_{i,j} = c + a_i$$

$$m_{i,j} = c * a_i$$

donde *c* es una constante y a_i es el ajuste para la fila *i*.

De modo similar, para el caso de valores constantes sobre columnas los biagrupamientos respetan alguna de las siguientes ecuaciones:

$$m_{i,j} = c + b_i$$

$$m_{i,j} = c * b_j$$

donde ahora c es una constante y bj es el ajuste para la columna j.

3. Biagrupamientos con valores coherentes

   En estos enfoques más sofisticados se indaga por biagrupamientos coherentes en ambas direcciones, en donde los valores de cada celda se generan por operaciones de adición o multiplicación. Por ejemplo, en el caso de la figura 11(d) cada celda de la segunda columna se pueden generar sumando 1 a las celdas de la primera columna, y las celdas de la tercer columna se pueden generar sumando 3 a las celdas de la segunda columna. Además los valores de la segunda fila se pueden generar sumando 1 a los valores de la primer fila, y los valores de la tercer fila se pueden generar sumando 2 a los de la segunda fila. De misma manera, en la figura 11(e) se observa un efecto similar pero con operaciones de multiplicación. En este caso los valores del biagrupamiento respetan alguna de las siguientes ecuaciones:

$$m_{i,j} = c + a_i + b_j$$

$$m_{i,j} = c * a_i * b_j$$

$$m_{i,j} = c * a_i + b_j$$

$$m_{i,j} = c * b_j + a_i$$

donde nuevamente *c* es una constante, a_i es el ajuste para la fila *i* y b_j es el ajuste para la columna *j.*

4. Biagrupamientos con evolución coherente

   Existe por último este cuarto tipo de biagrupamiento, donde todas las filas (resp. columnas) inducen un orden lineal a través de un subconjunto de columnas (resp. filas).

De acuerdo a las propiedades específicas de cada problema, uno o más de de estos tipos de biagrupamientos son considerados, y se deben usar diferentes tipos de funciones de mérito para evaluar la calidad de los mismos.

1,0	1,0	1,0	1,0
1,0	1,0	1,0	1,0
1,0	1,0	1,0	1,0
1,0	1,0	1,0	1,0

(a)

1,0	2,0	3,0	4,0
1,0	2,0	3,0	4,0
1,0	2,0	3,0	4,0
1,0	2,0	3,0	4,0

(b)

1,0	2,0	0,5	1,5
2,0	4,0	1,0	3,0
4,0	8,0	2,0	6,0
3,0	6,0	1,5	4,5

(c)

1,0	1,0	1,0	1,0
2,0	2,0	2,0	2,0
3,0	3,0	3,0	3,0
4,0	4,0	4,0	4,0

(d)

1,0	2,0	5,0	0,0
2,0	3,0	6,0	1,0
4,0	5,0	8,0	3,0
5,0	6,0	9,0	4,0

(e)

Figura 11. Tipos de Biagrupamientos: (a) valores constantes (b) valores constantes sobre filas (c) valores constantes sobre columnas (d) y (e) valores coherentes, aditivo y multiplicativo

A continuación se mencionan algunos de los algoritmos de biagrupamientos incluidos en el paquete de R *biclust* **[KEI2011]** utilizado en el presente trabajo.

1.9.1. Algoritmo CC

Representa el problema de biagrupamiento como un problema de optimización, definiendo un puntaje para cada grupo candidato y utilizando heurísticas para resolver las restricciones del problema de optimización definido por esta función de puntuación. La heurística es un algoritmo conocido como goloso (greedy) relajado, que da preferencias a las submatrices más grandes. Los autores asumen que los grupos tienen valores constantes, más posibles efectos aditivos sobre filas y columnas. Por lo tanto, después de remover los promedios sobre filas, columnas y submatrices, el residuo debería quedar tan pequeño como sea posible **[TAN2004]**.

1.9.2. Algoritmo Plaid

Es un enfoque estadístico, en donde la idea básica es representar la matriz de genes y condiciones como una superposición de capas, correspondientes a biagrupamientos. Se piensan diferentes valores como diferentes colores y las líneas de colores horizontales y verticales en la matriz correspondiente a una capa dan al método su nombre[4]. El modelo asume que el nivel de las entradas en la matriz es la suma de un fondo uniforme (gris) y k biagrupamientos cada uno coloreando una submatriz particular [TAN2004].

1.9.3. Algoritmo Quest

Este método se basa en una representación para los datos de expresión génica mediante motivos conservados de expresión génica o xmotifs. Se conserva un nivel de expresión génica a través de un conjunto de muestras si el gen es expresado con la misma abundancia en todas las muestras. Un motivo conservado de expresión génica es un subconjunto de genes que se conserva simultáneamente a través de un subconjunto de las muestras. En

[4] *Plaid* significa tela escocesa

[MUR2003] se presenta una técnica computacional para descubrir grandes motivos de genes conservados que cubren todas las muestras en los datos.

1.9.4. Algoritmo Bimax

Este algoritmo encuentra subgrupos maximales en una matriz binaria, para lo cual asume dos posibles valores por cada gen y condición: 1 si el gen responde de manera diferencial para esa condición y 0 si no lo hace **[PRE2006]**.

A continuación se muestra el funcionamiento del algoritmo utilizando como ejemplo la figura 12, donde las celdas en color gris representan los unos, y las celdas en color blanco los ceros.

Primero elije una fila *i* conteniendo ceros y unos (esta debe existir debido a que si la matriz solo tuviera ceros no existiría ningún bigrupo, y si solo tuviera unos existiría un único bigrupo extendido a toda la matriz).

Se divide las columnas en dos grupos: C_u son aquellas para las que la fila *i* es 1 y C_v aquellas para las que las que la fila *i* es 0, reordenando posiblemente para dejar juntos los unos y los ceros.

Posteriormente se dividen las filas en tres grupos: G_u son aquellas para las que existen unos en C_u pero no en C_v, G_w son aquellas para las que existen unos tanto en C_u como en C_v, y G_v son aquellas para las que existen unos en C_v pero no en C_u. Nuevamente reordenando para que queden los juntos los distintos grupos.

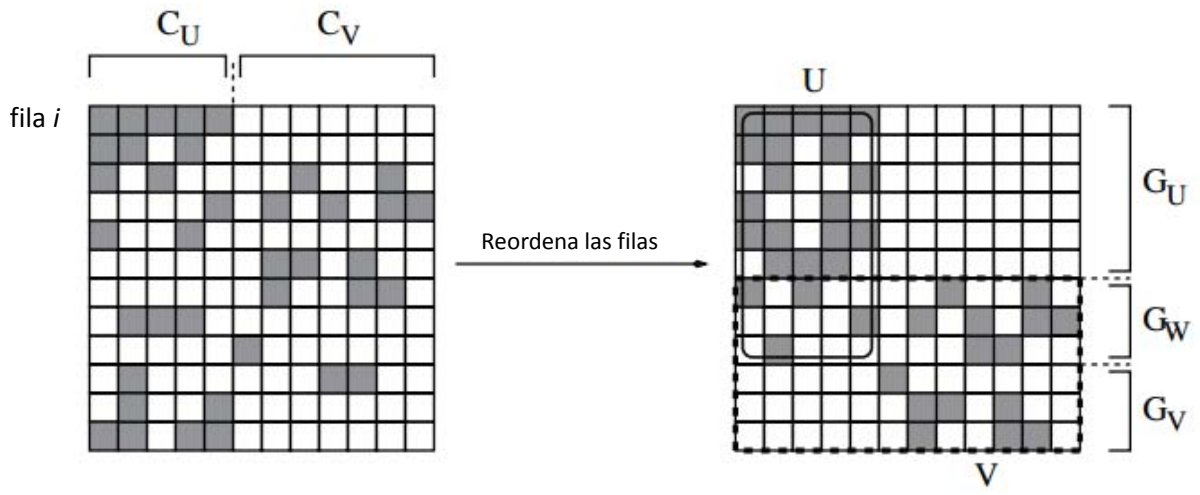


Figura 12. Algoritmo Bimax. Cada celda gris corresponde a un uno, y cada celda blanca a un cero.

Por último se realiza el mismo procedimiento para los conjuntos U (línea continua) y V (línea de trazos), deteniendo el proceso cuando se encuentre un conjunto formado por todos unos, esto es el bigrupo maximal.

1.9.5. Visualización de biagrupamientos

1.9.5.1. Visualización mediante mapas de calor

Una mapa de calor es una representación gráfica de los datos donde los valores contenidos en una matriz son representados como colores distintos, de manera que valores iguales son representados por el mismo color.

En la figura 13(a) se puede apreciar la misma muestra de los datos del experimento GSE6209 mencionada anteriormente. Como siempre, las columnas son los tratamientos y las filas los genes. A continuación, en la figura 13(b), su representación como mapa de calor, donde las columnas son los tratamientos, las filas los genes, y los valores de cada gen para cada tratamientos son mostrados en diferentes colores (gráfico realizado con el paquete de R *heatmap*).

	GSM143399	GSM143400	GSM143401	GSM143402
RV0001	0,967118	1,82321	1,11132	2,0513
RV0002	0,38549	1,46465	0,505222	-0,0613247
RV0003	0,189768	0,464284	-0,375436	-0,980485
RV0004	-0,420102	-0,253198	-0,164523	-0,672465
RV0005	0,350077	0,21535	-0,179201	0,271957
RV0006	-0,400041	-0,281939	0,507833	-0,234

Figura 13(a). Muestra de datos de la matriz construida a partir del experimento GSE6209.

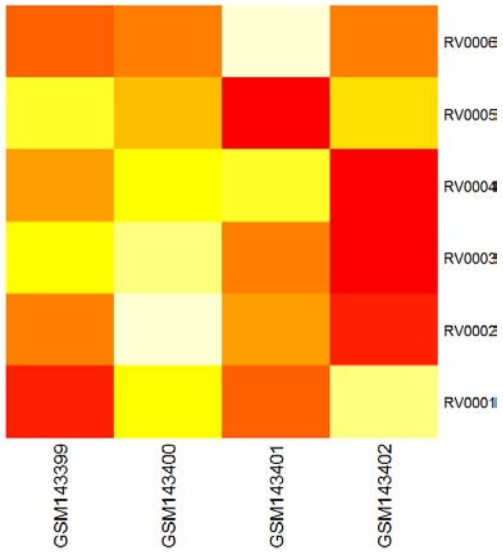


Figura 13(b). Mapa de calor para los datos correspondiente a la figura 13(a).

1.9.5.2. Visualización mediante gráficos de coordenadas paralelas

Las coordenadas paralelas constituyen un sistema de representación relativamente reciente. Su objetivo es resolver el problema de la

representación de conjuntos de datos multidimensionales. Se basan en representar cada dimensión como una escala vertical paralela a todas las demás. A cada elemento del conjunto de datos le corresponde una línea quebrada que une los valores que toman cada una de sus variables, el equivalente de un "punto" en representación bidimensional **[SII2009]**.

En la figura 14 se muestran los datos correspondientes al ejemplo anterior, ahora en un gráfico de coordenadas paralelas, donde el eje Y son los genes y cada una de las líneas representan los tratamientos (gráfico realizado con el paquete de R *MASS*). Cada línea une los valores que toma el tratamiento para los distintos genes.

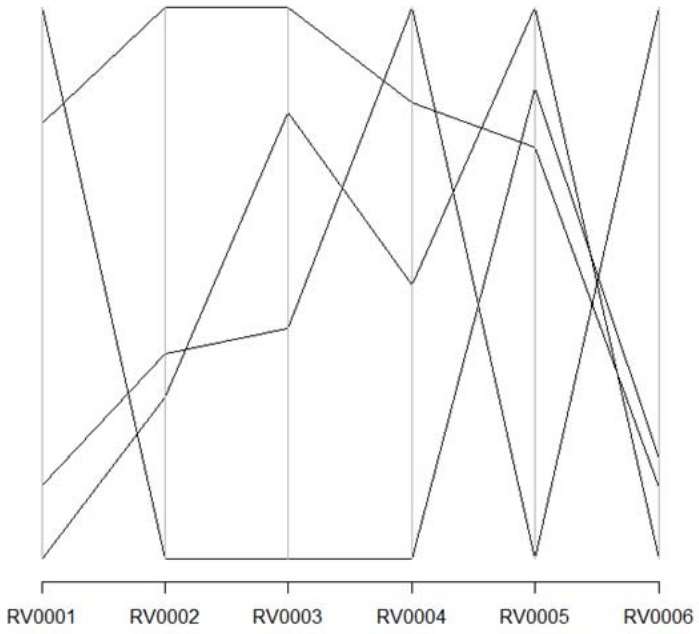


Figura 14. Gráfico de coordenadas paralelas correspondiente a los datos de la figura 13(a).

1.10. Validación de Agrupamientos

Dos de las más difíciles tareas en el análisis de agrupamientos son: Cómo decidir el número apropiado de grupos y cómo distinguir un mal agrupamiento de uno bueno. A continuación de describen formas de abordar estos problemas.

1.10.1. Silhoutte

Uno de los métodos utilizados para guiarnos en la respuesta a las preguntas anteriores es el Silhouette, donde a cada elemento se le asigna un valor s_i, que representa una medida de cuan bien el algoritmo lo ha asignado al grupo correcto. Cada grupo puede ser entonces, medido como un promedio de los s_i de cada uno de los elementos que lo constituyen, y finalmente el agrupamiento entero puede ser representado en forma gráfica, mostrando todos los valores s_i en un sólo diagrama, desde el cual se puede comparar la calidad de los grupos.

Consideremos un objeto i del conjunto de datos, y sea A el grupo al cual este objeto es asignado, entonces se calcula:

a (i) : Disimilitud promedio de i con todos los otros objetos de A

Ahora consideremos cualquier grupo C diferente de A y definamos

d (i, C) : promedio de disimilitud de i con todos los objetos de C

Computemos d (i, C) para todos los grupos C distintos de A y luego seleccionemos el más pequeño de ellos de la siguiente manera:

$b = \min d (i,C) , \quad C \neq A$

Sea B el grupo al cual pertenece este mínimo, y definamos b (i) = d (i,B).

Luego el valor s (i) se define como:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i),\ b(i)\}}$$

Se ve fácilmente que los valores de s(i) pertenecen al intervalo [-1, +1] y pueden ser interpretados como sigue:

S(i)	Interpretación
1	Objeto bien clasificado. La disimilitud con objetos dentro del grupo es mucho menor que la más pequeña disimilitud con objetos de grupos cercanos.
0	a(i) y b(i) son aproximadamente iguales. No está claro cuando i debería pertenecer a A o a B.
-1	Objeto mal clasificado. Su disimilitud con otros objetos dentro del grupo es mucho mayor que la disimilitud con objetos de grupos cercanos.

Tabla 3. Interpretación de la medida Silhouette según su valor

El *silhoutte* de un agrupamiento resume cuan apropiado es que un objeto pertenezca a un grupo determinado, y puede visualizarse como un gráfico con una línea horizontal cuya longitud es proporcional a cada s(i). El *silhoutte* muestra cuales objetos pertenecen adecuadamente a un grupo y cuales son objetos que deberían pertenecer a otro grupo o simplemente se encuentran entre dos grupos (figura 15).

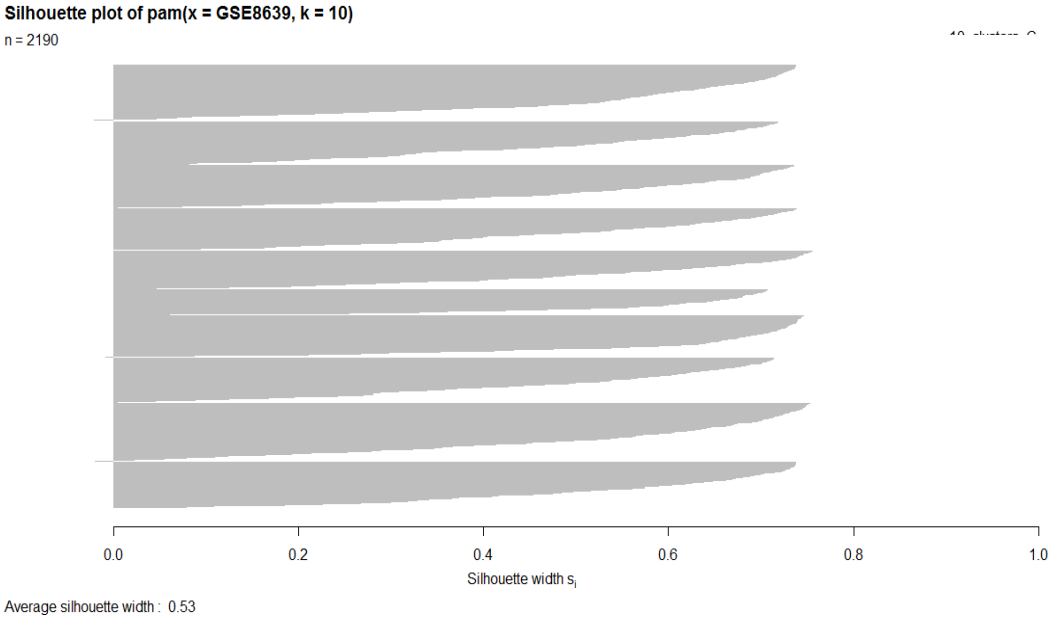


Figura 15. El gráfico de un Silhoutte para una argupamiento con k=10 utilizando una librería de R

El promedio de los s(i) para todos los elementos (a*verage silhouette width*) da un valor que representa una media de la calidad del agrupamiento completo.

El gráfico del *silhouette* es muy útil para decidir el número de agrupamientos. Se puede correr el algoritmo PAM muchas veces, cada vez para diferentes k y luego comparar el resultado del gráfico.

El promedio de los s(i) se puede usar para seleccionar el mejor número de grupos, eligiendo el k para el cual se obtiene el valor más alto.

$$SC = max_{\bar{S}}(k)$$

Donde máximo es tomado sobre todos lo k para los cuales el *silhoutte* puede ser construido. Este coeficiente es una medida de

la estructura del agrupamiento que ha sido descubierto y puede ser interpretado como sigue:

Rangos de SC	Interpretación
0.71-1.0	Estructura fuerte
0.51-0.70	Estructura razonable
0.26-0.50	La estructura es débil y podría ser artificial
0.25	Ninguna estructura substancial

Tabla 4. Interpretación de SC según su valor

1.10.2. Índice Rand Ajustado

El Índice Rand Ajustado (*Adjusted Rand Index*) es una medida utilizada para comparar la coincidencia entre dos agrupamientos (o la coincidencia entre un agrupamiento y un criterio externo). Este calcula la fracción de elementos correctamente clasificados sobre el total de elementos.

Supongamos que tenemos un conjunto de n objetos S = $\{O_1,\ldots, O_n\}$ y dos particiones de los objetos de C: $\{C_1,\ldots, C_k\}$; C': $\{C'_1,\ldots, C'_l\}$, siendo C el criterio externo y C' el resultado del agrupamiento sin pérdida de generalidad.

Además definamos los siguientes grupos de la manera siguiente:

S_{11} = {Pares de elementos que están en el mismo grupo tanto en C como en C'}

S_{00} = {Pares de elementos que están en diferentes grupos tanto C como en C'}

S_{10} = {Pares de elementos que están en el mismo grupo en C pero en diferentes en C'}

S_{01} = {Pares de elementos que están en diferentes grupos en C pero en el mismo en C'}

Donde $n_{ab} = |S_{ab}|$ para a,b ϵ {0, 1} y n es la cantidad de elementos del conjunto inicial.

Entonces el índice Rand se define como:

$$R(C, C') = \frac{2(n_{11} + n_{00})}{n(n-1)}$$

El rango de R va desde 0 si ningún par es clasificado en el mismo grupo bajo las dos agrupaciones, a 1 para agrupamientos idénticos.

Un problema del Índice Rand es que el valor esperado del Índice Rand de dos particiones aleatorias no toma un valor constante (digamos 0). Por lo tanto se propone un índice ajustado, el cual asume una distribución hipergeométrica generalizada como hipótesis nula: las dos agrupaciones se construyen aleatoriamente con un número de grupos fijos y un número fijo de elementos en cada grupo (el número de grupos en cada grupo no necesita ser el mismo). Si definimos *mij* como el número de elementos en la intersección de los grupos *Ci* y *C'j*:

$$m_{ij} = \left|C_i \cap C'_j\right| \qquad 1 \le i \le k, 1 \le j \le l$$

Luego el índice Rand ajustado es la diferencia del índice Rand y su valor esperado bajo la hipótesis nula y es definido como **[KUN2004]:**

$$R_{adj}(C, C') = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}\binom{m_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}$$

$$t_1 = \sum_{i=1}^{k} \binom{|C_i|}{2},\ t_2 = \sum_{j=1}^{l} \binom{|C'_j|}{2},\ t_3 = \frac{2t_1 t_2}{n(n-1)}$$

El índice así construido tiene el valor esperado en 0 para agrupaciones independientes y valor máximo 1 para agrupaciones idénticas.

1.11. Ontologías

Los sistemas de clasificación juegan un papel crucial en las actividades del hombre. En las ciencias, y en particular en las ciencias biomédicas, los investigadores describen los diferentes fenómenos mediante el uso de categorías y sub-categorías. Una ontología, sin embargo, es algo más que un sistema de clasificación; en una ontología, cada una de estas categorías, clases, términos o conceptos quedan definidos por una serie de aserciones que los conectan a otros términos.

El término ontología en ciencias de la información hace referencia a la formulación de un esquema conceptual exhaustivo y riguroso dentro de uno o varios dominios dados, con la finalidad de facilitar la comunicación y el intercambio de información entre diferentes sistemas o entidades. Aunque toma su nombre por analogía, ésta es la diferencia con el punto de vista filosófico de la palabra.

La ontología de genes (Gene Ontology o GO[5]), es un vocabulario controlado que representa el conocimiento acerca de atributos funcionales de los genes de una manera estructurada, y se puede usar en análisis tanto humanos como computacionales. Es un esfuerzo de colaboración para abordar la necesidad de descripciones coherentes de los productos de genes en diferentes bases de datos [VIE2010].

[5] Referirse al sitio http://www.geneontology.org/ para mayor información sobre GO

El proyecto GO ha desarrollado tres vocabularios estructurados controlados (ontologías) que describen los productos de los genes de una manera que es independiente de las especies (figura 16):

1. Procesos biológicos asociados (BP). Términos que describen una serie de funciones que concluyen en un objetivo biológico.
2. Componentes celulares (CC): Términos que describen en qué lugar de la célula el producto del gen está localizado.
3. Funciones moleculares (MF). Términos que describen la actividad bioquímica del producto de un gen.

Por otro lado hay tres aspectos distintos en este esfuerzo: en primer lugar, el desarrollo y el mantenimiento de las propias ontologías, en segundo lugar, la anotación de los productos de los genes, lo que implica establecer relaciones entre las ontologías y los genes y sus productos en las bases de datos, y tercero, el desarrollo de herramientas que faciliten la creación, el mantenimiento y el uso de ontologías.

El uso de los términos GO por las bases de datos colaborativas, facilita una consulta uniforme a través de ellas. Además los vocabularios controlados se estructuran de manera que se puedan consultar en diferentes niveles de abstracción. En la figura 17 se puede apreciar un detalle de una de las entradas de la ontología de procesos biológicos utilizando la herramienta AmiGo[6] para su visualización.

[6] http://amigo.geneontology.org/cgi-bin/amigo/go.cgi

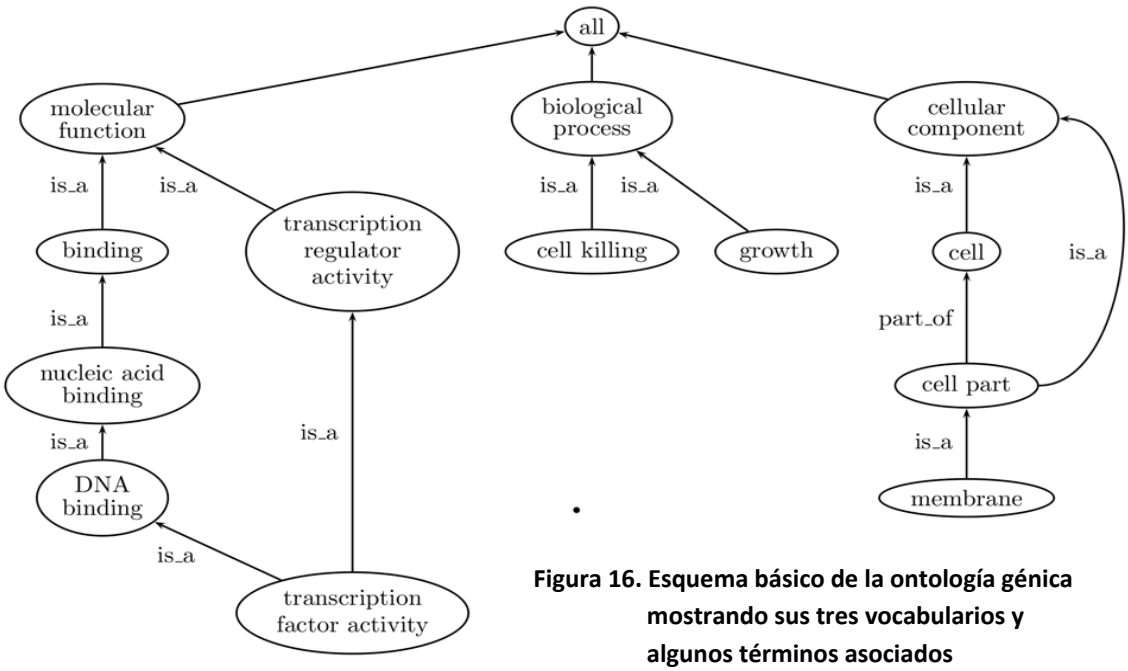


Figura 16. Esquema básico de la ontología génica mostrando sus tres vocabularios y algunos términos asociados

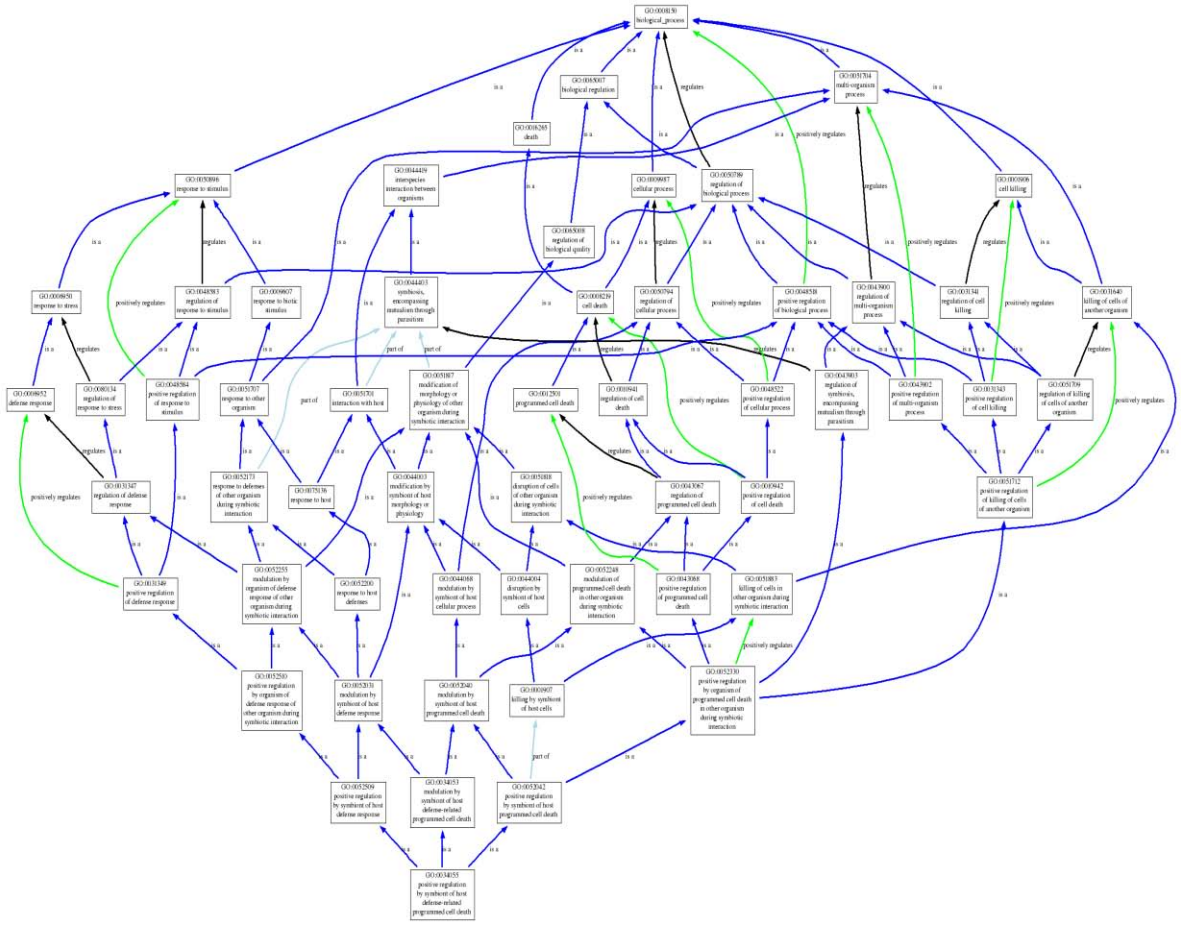


Figura 17. Visualización de una parte de la ontología génica, utilizando la herramienta AmiGO, donde se aprecia la complejidad de la misma

1.12. Similitud semántica basada en términos GO

Las medidas de similitud semántica permiten obtener valores numéricos en función de la cercanía del significado entre términos dados. En el caso de medidas de similitud basadas en ontologías, el significado se refiere a la anotación que tiene el término dentro de la ontología. En particular, la aplicación de las medidas de similitud semántica entre las anotaciones de GO proporciona una medida de su similitud funcional. En la actualidad, están disponibles diversas propuestas para cuantificar dicha similitud.

En los últimos años, las ontologías se han convertido en un tema principal en la investigación biomédica. Cuando las entidades biológicas se describen utilizando un esquema común, tales como una ontología, pueden ser comparadas por medio de sus anotaciones. Este tipo de comparación se denomina similitud semántica, ya que evalúa el grado de relación entre dos entidades por la similitud en el significado de sus anotaciones. Aunque la aplicación de la similitud semántica de ontologías biomédicas es reciente, en los últimos años se han publicado varios estudios, describiendo y evaluando los distintos enfoques. La similitud semántica se ha convertido en una valiosa herramienta para la validación de los resultados obtenidos de los estudios biomédicos, tales como la agrupación de genes, análisis de expresión génica de datos, predicción y la validación de las interacciones moleculares.

Los productos génicos se pueden anotar con muchos términos de GO de cada una de las tres ontologías; por tanto, la evaluación de la similitud funcional (dentro de una categoría particular de GO) implica comparar conjuntos de

términos en lugar de términos independientes. Un grupo de métodos determina la similitud funcional entre dos productos génicos a partir de la similitud semántica entre los términos a los cuales están anotados. En algunos casos se determina cada combinación de pares de términos anotados, mientras que en otros casos solo se consideran las mejores combinaciones.

Existen variantes que no se basan en la combinación de similitudes entre términos individuales para calcular la similitud semántica entre productos génicos, sino que la calculan directamente, por ejemplo, a partir de las anotaciones directas (no las heredadas). Estas metodologías son muy poco comunes, pues muy pocas medidas de similitud semántica consideran solo anotaciones directas. Otros métodos representan los productos génicos como los subgrafos de GO correspondientes a todas sus anotaciones (directas y heredadas). La similitud funcional se calcula utilizando técnicas de alineamiento de grafos o considerando los subgrafos como conjuntos de términos y utilizan alguna medida de similitud semántica. Las metodologías vectoriales representan el producto génico como un espacio vectorial, donde cada término corresponde a una dimensión; la similitud semántica se determina a partir de medidas de similitud vectorial **[IVE2010].**

Otro enfoque define una medida de similitud semántica basada sobre el contenido de información de ancestro común más informativo [RES1999]. El contenido de información de un concepto es inversamente proporcional a su frecuencia en el dominio (corpora). Los conceptos que son frecuentes en el dominio tienen un contenido de información bajo.

En **[JIA1997]** se propone una medida de distancia semántica basada sobre la diferencia entre los contenidos de información de los conceptos y el contenido de información del ancestro común más informativo. En [Lin] en cambio se propone una medida de similitud semántica basada sobre el proporción entre el contenido de información del ancestro común más informativo y l contenido de información de ambos conceptos.

Detallaremos a continuación una medida de similitud semántica basada en alineamientos de grafos conocida como Term Overlap, la cual fue implementada en R en el presente trabajo.

1.12.1. Term Overlap

En el cálculo de TO entre dos genes se considera para cada gen el conjunto de todas las anotaciones directas y todos sus términos ancestros asociados (excluyendo la raíz de la jerarquía), como un conjunto de anotaciones. Luego se calcula el puntaje TO para dos genes como el número de términos en la intersección de los dos conjuntos de anotaciones.

$$sim_{TO}(g_1, g_2) = \left| annot_{g_1} \bigcap annot_{g_2} \right|$$

Como en otras medidas, cuanto más alto es el puntaje mayor es la similitud entre los dos genes. El menor TO es cero y no existe un límite superior. Una variante es la forma normalizada (NTO), en la cual el TO es dividido por el

tamaño del conjunto de anotaciones para el gen con el número menor de anotaciones GO **[MEE2008]**.

$$sim_{NTO}(g_1, g_2) = \frac{\left|annot_{g_1} \cap annot_{g_2}\right|}{min\left(\left|annot_{g_1}\right|, \left|annot_{g_2}\right|\right)}$$

Finalmente la similitud para un conjunto G de genes, es la suma de la similitud de cada par de genes dentro del grupo, dividido por la cantidad de elementos del conjunto.

$$sim_{TO}(G) = \frac{\sum_{i,j}(g_1, g_2)}{|G|}$$

1.13. Blast y Blast2GO

BLAST (Basic Local Alignment Search Tool) es un programa informático de alineamiento de secuencias de tipo local, ya sea de ADN o de proteínas. El programa es capaz de comparar una secuencia de entrada contra una gran cantidad de secuencias que se encuentren en una base de datos. El algoritmo encuentra las secuencias de la base de datos que tienen mayor parecido a la secuencia de entrada. Es importante mencionar que BLAST usa una heurística, por lo que no nos puede garantizar que ha encontrado la solución óptima. Sin embargo, BLAST es capaz de calcular la significación de sus resultados, por lo que nos provee un parámetro para juzgarlos.

Uno de los usos comunes de BLAST es para encontrar probables genes homólogos. Por lo general, cuando se obtiene una nueva secuencia, se usa

BLAST para compararla con otras secuencias que han sido previamente caracterizadas, para así poder inferir su función. BLAST es la herramienta más usada para la anotación y predicción funcional de genes o secuencias proteicas. Se han creado muchas variantes para resolver algunos problemas específicos de búsqueda.

Blast2GO es una herramienta bioinformática para anotación funcional y análisis de genes o proteínas. La herramienta se desarrolló originalmente para proveer una interfaz amigable a las anotaciones de la ontología génica, pero las nuevas versiones cuentan con funcionalidades nuevas.

Básicamente Blast2GO usa una búsqueda BLAST para encontrar secuencias similares a una o varias secuencias de entrada para las cuales se desconocen los términos GO. Luego el programa extrae los términos GO asociados a cada una de las secuencias obtenidas por BLAST. De esta manera se le pueden asignar términos GO a las secuencias de entrada.

1.14. *Reconocimiento de patrones*

Dadas dos cadenas c_1 y c_2 de caracteres (que pueden representar cadenas de ADN), el reconocimiento de patrones, en el contexto del trabajo presente, intenta encontrar subcadenas s_1 perteneciente a c_1 y s_2 perteneciente a c_2, que sean “similares” entre sí. El grado de similitud implica desde una correspondencia perfecta de uno a uno entre sus caracteres, a algún grado de disimilitud permitido. Se asigna una puntuación (score) de acuerdo a qué

tan parecidas sean estas subcadenas. Este concepto se puede extender a la búsqueda de patrones entre un número N de cadenas.

En particular se quiere identificar de manera automática, patrones dentro de una colección de secuencias, correspondientes a regiones promotoras de genes pertenecientes a un mismo grupo dentro de un agrupamiento dado. Es de esperar que si los genes pertenecen al mismo grupo, y por lo tanto se expresan de manera similar ante las distintas condiciones de los experimentos (o al menos dentro de algunas de estas condiciones), existan en la regiones promotoras uno o más motivos (motifs) que puedan explicar este comportamiento.

En [MAH2004] se propone una solución basada en mapas auto organizados, y en [BAI1994] otro basado en EM (Expectation Maximization) que se detalla a continuación.

1.14.1. Meme y MAST

MEME (*Mutliple EM for Motif Elicitation*) provee un portal[7] unificado para descubrir y analizar motivos que representan características tales como sitios de unión (binding site) en el ADN.

En [BAI1994] se describe un algoritmo para descubrir uno o más motivos en una colección de ADN o secuencia de proteínas utilizando la técnica de EM para ajustar un modelo finito de dos componentes al conjunto de secuencias. Para encontrar múltiples motivos se ajusta el modelo a los datos, se borra

[7] Referirse a http://meme.nbcr.net/meme/ para mayor información sobre Meme y Mast

probabilísticamente las ocurrencias de los motivos que se han encontrado, y se repite el proceso para encontrar nuevos motivos. El algoritmo requiere solamente un conjunto de secuencias sin alinear y un número especificando el ancho de los motivos. Este retorna un modelo para cada motivo.

MEME espera como entrada un archivo en formato FASTA. Como mencionamos en la introducción, este es un formato basado en texto utilizado para representar secuencia de ADN y en el que los nucleótidos se representan con la letra correspondiente (A, C, G o T). En esta secuencia colocamos una entrada con la denominación del gen (por ejemplo Rv0165c), y debajo la secuencia de la región intergénica, como puede observarse en el siguiente ejemplo:

```
>Rv0165c
CTATCCCTTCCTTCCCTTCCTTTCACTTGCGCATCCCTTGCGCCAGGTTGATAT
>Rv0258c
CGCCACGCCCGAAGCCACAGAGGTGGGTATCGGCAATGGGCAATCCGGCAGCAATTGCCTGG
TAACGCGACTGAAACCTCACAGGCCTAGACACGTCAT
>Rv0282
CGGCGCACCGTTCGCGCTGCCGGCACCCCGGGCTCCATAATGAAAATCATGTTCAGTAAGCTA
CACTCTGCATATCGGGCTACCAACGAAATGGAGTATCGGTCATGATCTTGCCAGCCGTGCCTAA
AAGCTTGGCCGCAGGGCCGAGTCGATTGGTCGCGGTCGCCTCGACAGTTAGCTTATGCAATGC
TAACTTCGGGGCAAAGTTCAGGCGGATCGGCCG
```

Una vez ejecutado se podrán visualizar los patrones encontrados por el algoritmo y los sitios de ocurrencia de dicho patrón (se pueden ver ejemplos de la salida en la segunda parte de la tesis).

Por otro lado, MAST (Motif Alignment & Search Tool) permite buscar un patrón determinado en todas la regiones intergénicas. De esta manera podemos ver si el motivo que ha descubierto MEME existe en algún otro sitio. Si se encuentra un motivo en un grupo de genes, resulta interesante determinar si ese motivo está relacionado con otros genes, dado que no necesariamente estos genes deben pertenecer al grupo que se está analizando.

1.15. Descripción del presente trabajo

Uno de los usos interesantes de experimentos con microarreglos es la búsqueda de agrupamientos de genes que comparten el mismo perfil de expresión, debido a que ello puede revelar cómo el metabolismo responde al entorno. Estos agrupamientos se pueden encontrar utilizando algunos de los métodos estadísticos de agrupamiento convencionales (PAM, CLARA y HOPACH están entre los más mencionados en la literatura para trabajar con microarreglos) o utilizando los métodos más novedosos conocidos como biagrupamientos. En ambos casos los grupos obtenidos se deben evaluar con las medidas de significancia correspondiente a cada método, con el fin de determinar la validez de los grupos obtenidos.

Por otro lado, y con el fin de determinar la validez biológica de los grupos encontrados, se hace necesaria una validación semántica. Para esta validación se utilizará una medida de similitud semántica conocida como superposición de términos (term overlap o TO) sobre las anotaciones de los

términos GO (Ontología Génica o Gene Ontology), determinando una validez biológica de los grupos desde el punto de vista de dicha ontología.

Una vez validados estadística y semánticamente, para cada grupo de genes se recupera la región intergénica de sus miembros (se verá en la segunda parte de la tesis cómo se recuperaron las regiones intergénicas), y se genera una secuencia en formato FASTA.

Esta secuencia se presenta a los algoritmos de búsqueda de patrones con el fin de determinar la existencia de posibles sitios regulatorios comunes a todo el grupo. Por sitio regulatorio se entiende una zona de la región intergénica que regulan la activación del gen, y es de esperar que genes hallados en un mismo grupo, y que por lo tanto se expresan de manera similar, tengan algún sitio regulatorio común que pueda explicar este comportamiento. Para la búsqueda de estos patrones se utiliza MEME (*Multiple EM for Motif Elicitation*).

Para el estudio de los experimentos se diseñó un flujo de procesos, el cual cubre cada uno de los puntos teóricos explicados en la primera parte. Muchos de estos procesos fueron desarrollados en R, utilizando las librerías necesarias en cada caso. Así se desarrollaron programas en R para la obtención de los datos desde las bases del NCBI, el preprocesamiento, los distintos algoritmos de agrupamientos y biagrupamientos, y los algoritmos de similitud semántica. En la figura 18 se presenta una vista general de este flujo.

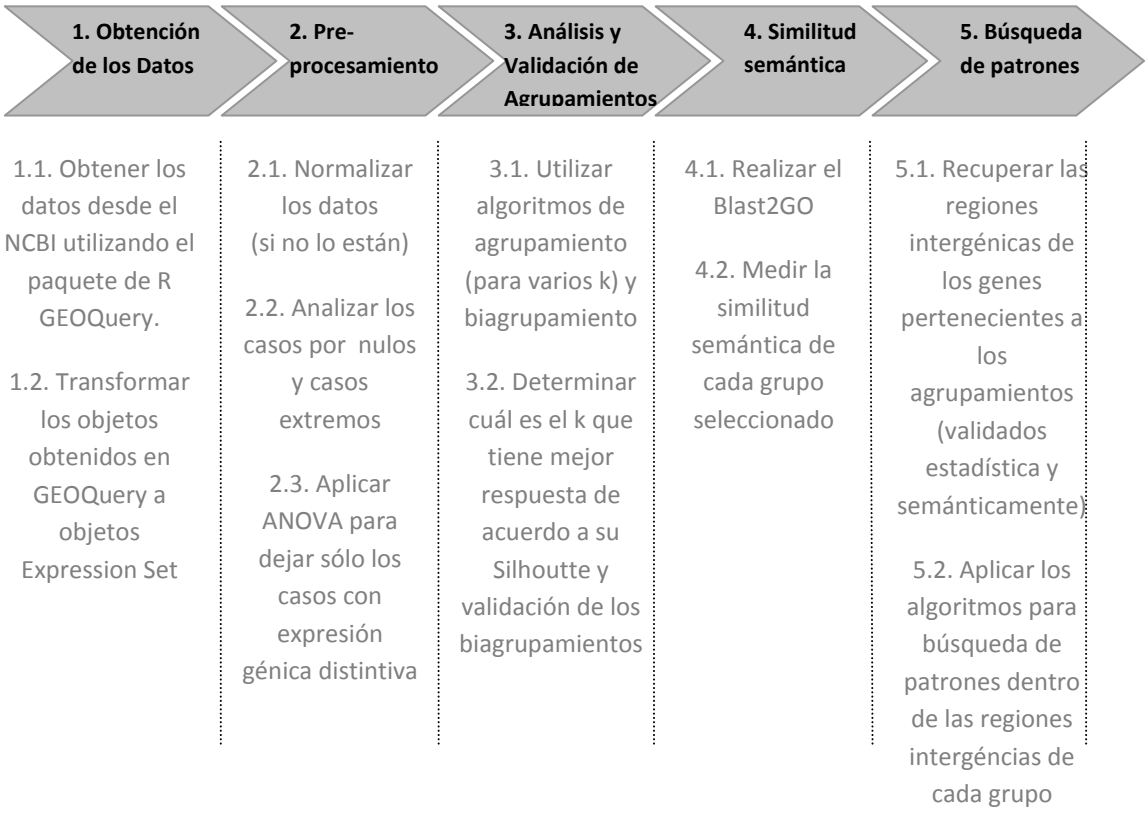


Figura 18. Flujo de procesos

2. Desarrollo

2.1. Obtener los datos desde el NCBI

En este paso se obtienen los datos desde el NCBI utilizando el paquete GEOQuery de R. GEOQuery es un conjunto de funciones para recuperar la información desde la base pública del NCBI Gene Expression Omnibus (GEO), proporcionando una interfaz de esta base hacia la plataforma bioconductor. La función *getGEO* es la encargada de bajar la información en formato SOFT (compatible con MIAME), devolviendo una estructura desde donde se puede explorar toda la información referente al experimento.

Reiteramos aquí la tabla conteniendo los conjuntos de datos a utilizar:

Tabla de experimentos seleccionados		
Identificador	Descripción	**# muestras**
GDS2677	**Efecto de la capreomicina sobre Mycobacterium tuberculosis**	4
GSE6209	**Mycobacterium tuberculosis y respuesta a macrófagos**	11
GSE8639	**Regulación del gen que codifica la alfa-cristalina de Mycobacterium tuberculosis en la respuesta hipóxica**	6
GSE10391	**Dormición In vitro lograda por multiples estreses en Mycobacterium tuberculosis**	75
GSE12364	**Respuesta trancripcional global a la vancomincina en Mycobacterium tuberculosis**	12
GSE15976	**Respuesta del factor sigma B (sigB) a condiciones de estrés (SDS 0.05% y diamida 5mM)**	36
GSE365	**GSE365 Reparación del ADN en Mycobacterium tuberculosis**	28

GSE7962	**El factor sigma E de Mycobacterium tuberculosiscontrola la expression de componentes bacterianos que modulan macrófagos**	23
GSE9776	**Efecto del tratamiento con INH en la expression génica de Mycobacterium tuberculosis en varios modelos de dormición**	17

Tabla 5. Experimentos de microarreglos seleccionados, donde se observa el identificador del experimento, la descripción y el número de muestras.

Luego de recuperados los datos desde el NCBI, y como se mencionó anteriormente, para cada experimento se construye una matriz en donde sus filas corresponden a los genes y sus columnas a los distintos tratamientos del experimento. Sobre cada una de estas matrices se realizaron los agrupamientos y biagrupamientos de manera independiente.

2.2. Preprocesamiento

Una vez recuperado el experimento a través de la librería GEOquery, hay algunas consideraciones a tener en cuenta antes de proceder a realizar el paso de preprocesamiento. Estas consideraciones pueden requerir de alguna acción previa, con el fin de evitar inconsistencias posteriores.

a. Algunos experimentos hacen referencia a distintas plataformas. Es conveniente para evitar inconsistencias seleccionar los tratamientos correspondientes a una de las plataformas al momento de analizar los datos.

b. Los experimentos pueden tener distinto tipo de normalización de sus resultados. Dentro de la información referente a los experimentos existe un atributo VALUE el cual indica qué tipo de transformación se ha realizado a las lecturas del microarreglo (por ejemplo, en el experimento GSE10391 a este atributo le corresponde el valor *Print Tip Lowess Normalized Log 2 ratio test/reference*, que indica que el resultado es el valor normalizado de la manera indicada).

c. Tener precaución de verificar la referencia de cada canal, dado que en algunos casos el tratamiento puede corresponder a un canal, y en otros corresponder a otro canal.

Una vez comprendido cómo se están mostrando los valores dentro de cada experimento, pasamos a realizar distintos preprocesamientos al mismo. Como primer paso nos quedamos con los genes cuyos nombres comiencen con RV que son los de nuestro interés, por ejemplo eliminando de nuestro análisis los controles propios del microarray.

Como ejemplo de este primer paso veamos como es el tratamiento para el experimento GSE10391. Este experimento tiene 75 muestras provenientes de dos plataformas:

Plataformas del experimento GSE10391		
Identificador	Descripción	**# entradas**
GPL4369	**UCF BMS M. tuberculosis v1.0**	10800
GPL4388	**UCF BMS M. tuberculosis ver 2.0**	22080

Tabla 6. Plataformas del experimento GSE10391

La primera con 10800 observaciones y la segunda con 22080. Por uniformidad del análisis nos quedamos con la primera de ellas.

Como siguiente paso se analizan los valores faltantes que puedan contener nuestras muestras. De acuerdo al algoritmo de análisis que apliquemos para el análisis, estos valores pueden necesitar de un tratamiento particular. Por ejemplo, en el caso de biagrupamiento los algoritmos requieren que no

existan valores faltantes, para lo cual se aplicará un paso de imputación. Se utiliza el paquete impute de R [HAS2011] que dispone diferentes funciones para completar estos valores nulos. El utilizado en el trabajo presente es la función impute.knn, que utiliza un algoritmo de vecino más cercano utilizando una métrica euclídea para imputar valores faltantes. En el caso en el que los valores faltantes constituyan más del 50 % de la variable, se procederá a desestimar dicha variable.

Posteriormente se desarrollaron una serie de filtros sencillos ad-hoc con el fin de reducir la cantidad de genes de nuestra muestra original y facilitar el funcionamiento de los algoritmos de agrupamiento. Algunos de estos filtros son:

Retener solos los genes RV: Este filtro se aplica siempre, debido a que son los genes que buscamos analizar, porque corresponden a los genes de la cepa H37rv de *Mycobacterium tuberculosis*, que se utiliza ampliamente como modelo.

Excluir genes mayores: Elimina los genes en donde todos sus valores supera un umbral predeterminado. Tiene dos parámetros, uno es el valor a exceder y cuyo valor por defecto es 100 y otro es la proporción de muestras que cumplen esta condición para que el filtro sea aplicado, con un valor por defecto es 0.05.

Excluir genes menores: Elimina los genes donde todos sus valores están por debajo de un umbral predeterminado. Similar al anterior, posee dos parámetros: uno es el valor por debajo del cual el gen es excluido, cuyo valor por defecto es 0.5, y otro es la proporción de muestras que se espera cumplan con esta condición, también para que el gen sea excluido, con un valor por defecto de 0.05.

Excluir genes nulos: Elimina los genes que tienen una proporción de nulos mayor a un umbral determinado. Si el valor está por debajo del umbral es posible aplicar una función para completar el valor faltante. El valor por defecto de este filtro es 0.1.

Excluir homogéneos: Elimina los genes cuyos valores están todos dentro de un rango especificado para todas las muestras. Para cada gen se toma la diferencia absoluta entre su valor máximo (el valor de la muestra que tiene el valor más alto para ese gen) y su valor mínimo (el valor de la muestra que tiene el valor más bajo para ese gen) y se excluye si esta diferencia es menor a un valor A predeterminado. El valor por defecto de A es 0.1.

Normalizar: Normaliza la muestra (utilizando la librería marray **[YAN2012]).**

Estos filtros fueron probados con la expectativa de reducir el número de observaciones, intentando dejar solamente aquellas que puedan ser útiles al análisis, sin demasiado éxito, motivo por el cual finalmente se utilizaron filtros más complejos, disponibles en la librería de R genefilter. Esta librería tiene una serie de filtros a aplicar sobre el conjunto de genes, los cuales pueden ser especificados dentro de la función genefilter. Algunos de estos filtros son:

Anova: Devuelve una función que al ser evaluada sobre el conjunto de genes realiza un Anova, que retorna TRUE si el valor p para una diferencia entre medias es menor a un p establecido. El valor por defecto de p es 0.05.

kOverA: Devuelve una función que evalúa a TRUE si al menos k de los elementos son mayores que un A establecido. El valor por defecto de A es 100.

pOverA: Devuelve una función que evalúa a TRUE si la proporción de valores mayores a A es al menos p. El valor por defecto de A es 100 y el de p es 0.05.

cv: Devuelve una función que computa el coeficiente de variación para el vector de entrada y retorna TRUE si este valor está entre los valores establecidos a y b. El valor por defecto de a es 1 y el de b es 100.

La aplicación de este filtro requiere cargar los datos del experimento en una clase ExpressionSet, la cual permite combinar diferentes fuentes de información del microarreglo en una estructura única y conveniente **[FAL2006].**

Consideremos la aplicación de *Anova* como filtro. En este filtro los factores son cada uno de los tratamientos. Para indicar esto primero se construye un vector de la misma longitud que la cantidad de muestras, identificando con la misma letra a los tratamientos que son réplicas de las mismas condiciones y por letras distintas a tratamientos con diferentes condiciones. Por ejemplo, veamos el caso del experimento GSE6209, donde este vector toma el siguiente aspecto: [A A A A A B B B B B A], dado que las 5 primeras muestras y la última son réplicas del experimento bajo las mismas condiciones, y las muestras de la 6 a la 10 son réplicas del experimento bajo otras condiciones. Por otro lado para la aplicación del *Anova* se utilizó un *p-value* de 0.05.

Muestras del experimento GSE6209	
Muestra	**Descripción**
GSM143399	**H37Rv 4 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 1 (A 4 horas)**

GSM143400	**H37Rv 4 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 2 (B 4 horas)**
GSM143401	**H37Rv 4 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 3 (C 4 horas)**
GSM143402	**H37Rv 4 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 4 (D 4 horas)**
GSM143403	**H37Rv 4 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 5 (E 4 horas)**
GSM143404	**H37Rv 24 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 1 (A 24 horas)**
GSM143405	**H37Rv 24 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 2 (B 24 horas)**
GSM143406	**H37Rv 24 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 3 (C 24 horas)**
GSM143407	**H37Rv 24 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 4 (D 24 horas)**
GSM143408	**H37Rv 24 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 5 (E 24 horas)**
GSM206719	**H37Rv 4 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 6 (F 4 horas)**

Tabla 7. Las 11 muestras del experimento GSE6209 con sus respectivas descripciones.

Llamemos al vector anterior T, luego podemos construir el filtro de la siguiente manera:

Anova(factor(T), p=0.01, na.rm=TRUE)

Aplicando este filtro al experimento GSE6209 mediante la función genefilter, se reduce el conjunto inicial de 3924 genes a solamente 162 (los que pasaron el test para el valor de p dado), visualizados en el diagrama de caja de las figuras 19(a) y 19(b) .

Debido a la gran cantidad de test, es posible encontrar algunos falsos positivos, pero el propósito del filtro es descargar datos con ruido para la

posterior aplicación de agrupamientos más que obtener una lista de genes diferencialmente expresados.

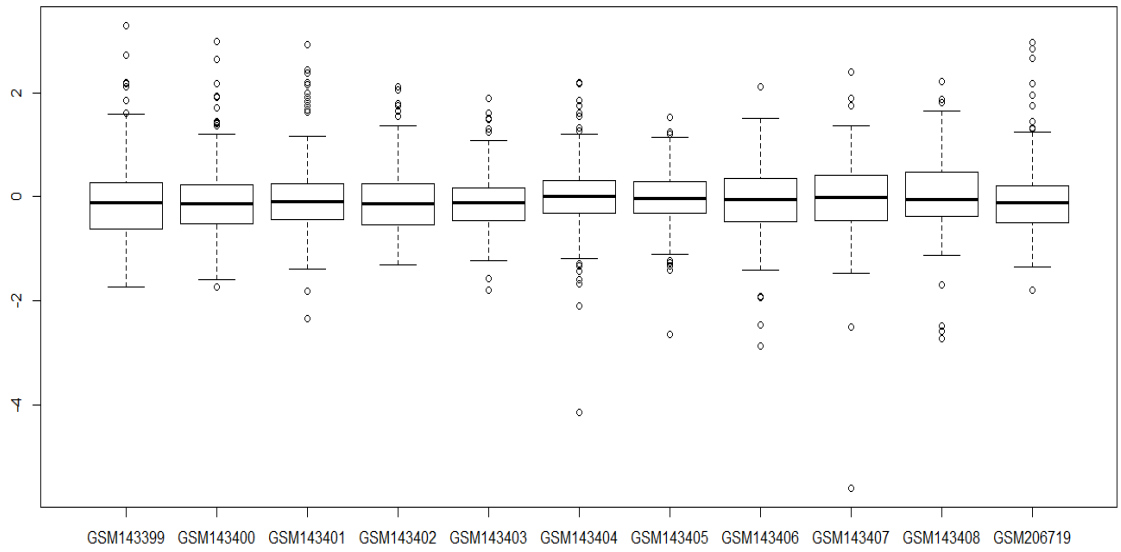


Figura 19(a). Diagrama de caja del experimento GSE6209 antes de aplicar el filtro Anova.

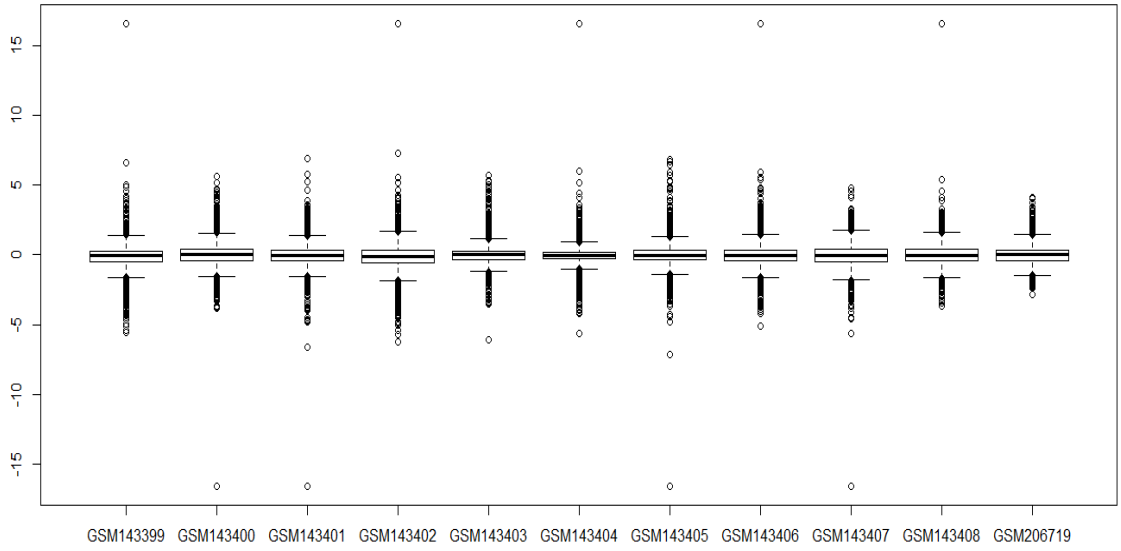


Figura 19(b). Diagrama de caja del experimento GSE6209 luego de aplicar el filtro Anova de la librería genefilter.

2.3. Agrupamientos y biagrupamientos

Se exploraron dos alternativas al momento de realizar los agrupamientos: algoritmos convencionales por un lado y biagrupamientos por el otro.

Para el caso de los algoritmos convencionales se probaron algunos en función de su adecuación para resolver problemas con microarreglos. Esto son PAM, CLARA y HOPACH (todos se pueden encontrar en el paquete de R cluster). Finalmente se seleccionaron los métodos PAM y CLARA por ser mediante los cuales se obtuvieron los mejores resultados (comparados con la medida Silhoutte). De todas maneras se mostrarán algunos resultados utilizando HOPACH, por ser este mencionado en la literatura como muy adecuado para agrupamientos con microarreglos.

Para el caso de los biagrupamientos se utilizaron los siguientes algoritmos: CC, Plaid, Bimax y Quest (todos ellos se pueden encontrar en la librería de R biclust).

2.3.1. Agrupamientos convencionales

En esta sección se detallan los procesos y resultados de agrupamientos obtenidos mediante la ejecución de los algoritmos convencionales, mostrando en cada caso su validación mediante el método s*ilhoutte*.

Para automatizar el procesamiento se realizó un programa en R, el cual toma como entrada uno de los experimentos, y ejecuta PAM y CLARA para

distintos k, seleccionando el más adecuado de acuerdo a la medida *silhouette*.

Como primera medida se realizaron los agrupamientos con los filtros básicos detallados en apartados precedentes. Los resultados obtenidos no cumplían totalmente con los requisitos necesarios para ser utilizados en el descubrimiento de sitios regulatorios potenciales, donde el principal requisito es encontrar al menos un grupo razonablemente pequeño (entre 20 y 80 genes) como para poder aplicar técnicas de búsqueda de patrones.

Detallamos a continuación los resultados para el algoritmo CLARA (tener presente que el algoritmo CLARA es una extensión de PAM para tratar con mayores volúmenes de datos utilizando una técnica de muestreo). Estos resultados se muestran en gráficos, y son producto de 100 corridas automáticas donde se va variando el valor de k y mostrando algunos indicadores para determinar el k más adecuado. Además de variar el k, se van variando distintas alternativas de filtrados de acuerdo a los filtros básicos vistos en el apartado anterior. Solo se mostrarán los resultados en donde los filtros aplicados dieron los mejores resultados.

En todos los casos se muestra el s*ilhouette width* para cada k y el s*ilhouette* completo para el k seleccionado. Se parte de un k mínimo de 100 debido a que para k menores aparecen generalmente todos grupos muy grandes como para poder encontrar patrones. La idea es encontrar algunos grupos pequeños (de hasta 50 genes), de tal manera que los algoritmo de búsqueda de patrones funcionen correctamente, dado que para tamaños mayores de genes el algoritmo encuentra siempre patrones que pueden ser ficticios. En

particular los grupos seleccionados serán grupos validados estadística y semánticamente (utilizando las medidas de *silhoutte* y *term overlap* respectivamente).

Cabe aclarar que no es de tanta utilidad encontrar agrupamientos con un buen s*ilhoutte* general, que mide el comportamiento general de toda la partición, que con grupos relativamente pequeños con buen *silhouette width*, que indican grupos interesantes a ser tenidos en cuenta para el análisis con la herramientas de búsqueda de patrones.

2.3.1.1. Agrupamientos utilizando el algoritmo CLARA

Como se mencionó anteriormente para cada experimento se realizaron 100 corridas de algoritmo CLARA, para distintos valores de k entre 100 y 200. La elección de este rango de valores es motivada por el hecho de querer encontrar al menos un grupo pequeño (20 a 80 genes). Para cada k se obtuvieron los valores de *silhouette width* indicando la bondad del agrupamiento, y se seleccionaron los dos mejores grupos de acuerdo a esta medida. En la figura 20 se puede observar un gráfico de los valores de silhoutte width en función del k, para el experimento GSE6209, en donde se marcan con círculos rojos los valores máximos seleccionados, correspondientes a k1=101 y k2=162. Para estos dos agrupamientos seleccionados, se muestran en la figura 21 sus silhoutte completos, en donde se aprecian buenos agrupamientos, con valores mayores a 0,8 para algunos grupos.

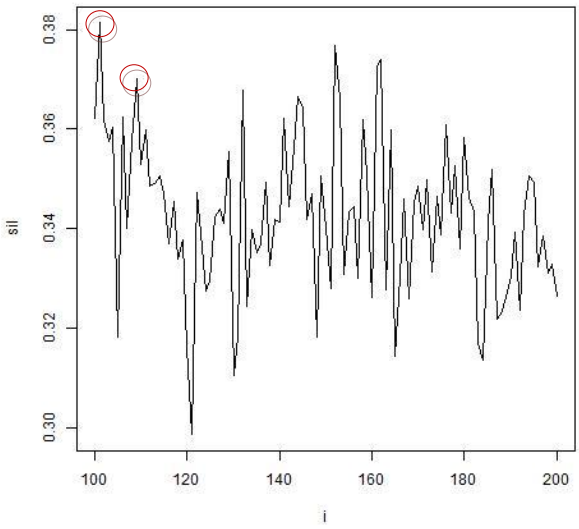
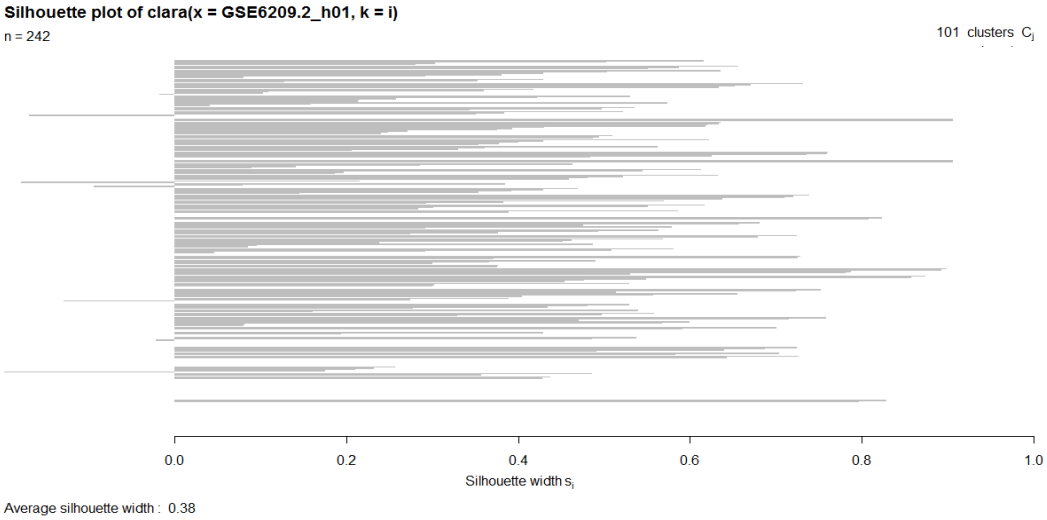


Figura 20. Silhoutte width en función de la cantidad k de particiones



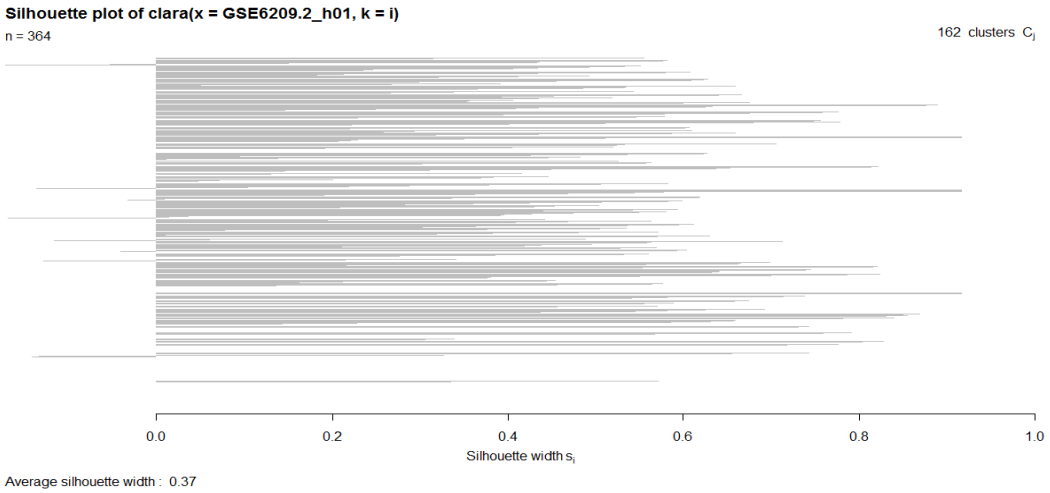


Figura 21. Silhoutte para k=101 y k=162

Este procedimiento se realizó para 4 experimentos (ver los resultados completos en el apéndice A), obteniéndose la siguiente tabla de resultados.

K seleccionados por experimento				
Experimento	K_1	**Silhoutte width$_1$**	K_2	**Silhoutte width$_2$**
GSE6209	101	0,38	162	0,37
GSE10391	100	0,23	157	0,21
GSE8639	110	0,52	120	0,50
GSE12364	142	0,25	183	0,25

Tabla 8. Número de grupos seleccionados al aplicar el algoritmo CLARA a cada experimento para los dos mejores Silhouette Width.

2.3.1.2. Agrupamientos utilizando el algoritmo HOPACH

Utilizando el algoritmo HOPACH (una técnica que tiene la ventaja de no requerir como entrada el número k de cantidad de grupos) para los experimentos detallados anteriormente, los resultados muestran un k de 2 o

3 grupos si se usan las distancias del coseno (que es la que viene por defecto) o la de correlación. Como ejemplo vemos el comportamiento de este algoritmo para el experimento GSE6209 posteriormente a ser aplicados los filtros básicos.

El resultado son dos grupos con 1665 y 1965 elementos respectivamente.

Sin embargo si se utiliza la distancia euclídea para este mismo caso se obtiene un agrupamiento compuesto de 80 grupos. Utilizando silhouette se puede apreciar que si bien ciertos elementos están mal agrupados, y hay otros grupos que por tener muchos elementos carecen de la condiciones necesarias para el análisis propuesto, hay otros que pueden resultar promisorios, en donde sería deseable continuar investigando, por ser grupos pequeños con un silhoutte razonable (los grupos que muestran un *silhouette width* mayor a 0,35). En la figura 22 se muestra en el recuadro rojo un grupo para el cual gran parte de los genes debería pertenecer a algún otro grupo según su *silhouette*, pero aún así se pueden recuperar otros grupos más pequeños de mejor calidad.

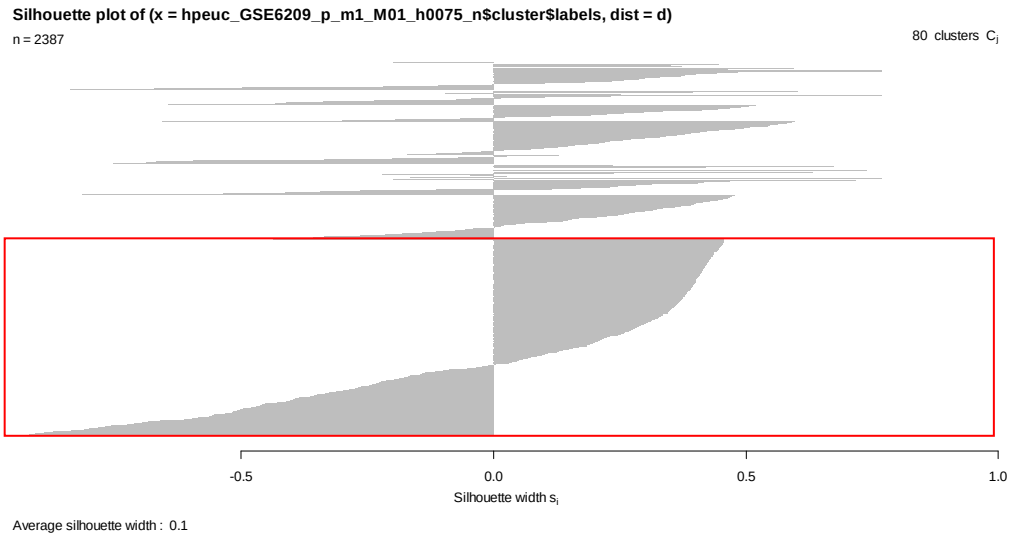


Figura 22. Silhoutte (80 grupos determinados por HOPACH)

Este procedimiento se realizó para 4 experimentos (ver los resultados completos en el apéndice B), obteniéndose la siguiente tabla.

K determinado por experimento		
Experimento	**K**	**Silhoutte width**
GSE6209	80	0,10
GSE10391	175	-0.03
GSE8639	298	0,15
GSE12364	9	0,15

Tabla 9. Número de grupos al aplicar el algoritmo HOPACH a cada experimento

2.3.2. Agrupamientos convencionales aplicando filtros de ANOVA

En este apartado se repiten los mismos procedimientos realizados anteriormente, pero esta vez sobre los experimentos filtrados mediante el filtro *Anova*. Para aplicar este filtro al conjunto de datos se utiliza la función Anova, provista dentro de la librería genefilter **[GEN2009].** Recordemos que Anova devuelve una función que al ser evaluada sobre el conjunto de genes realiza un Anova, que retorna TRUE si el valor p para una diferencia entre medias es menor a un p establecido.

2.3.2.1. Agrupamientos filtrados con ANOVA utilizando el algoritmo CLARA

Tomemos por ejemplo el experimento GSE6209. Primero, tal como se explicó en el apartado 3.1.2, se realiza el Anova sobre el conjunto de datos, dejando 162 observaciones (genes) de los 3924 originales. Posteriormente se realiza

un proceso de agrupamiento utilizando el algoritmo CLARA similar al que se aplicó anteriormente, resultando un k=16 como adecuado para este caso.

En la figura 23 se muestra el *silhoutte* obtenido, donde se aprecia una mejora a nivel global del agrupamiento (*silhoutte width* = 0.22). También se observan que se definen grupos pequeños con un buen *silhoutte* (mayores a 0.5), adecuados para un posterior análisis.

El detalle con el análisis para el resto de los experimentos se muestra en el apéndice C.

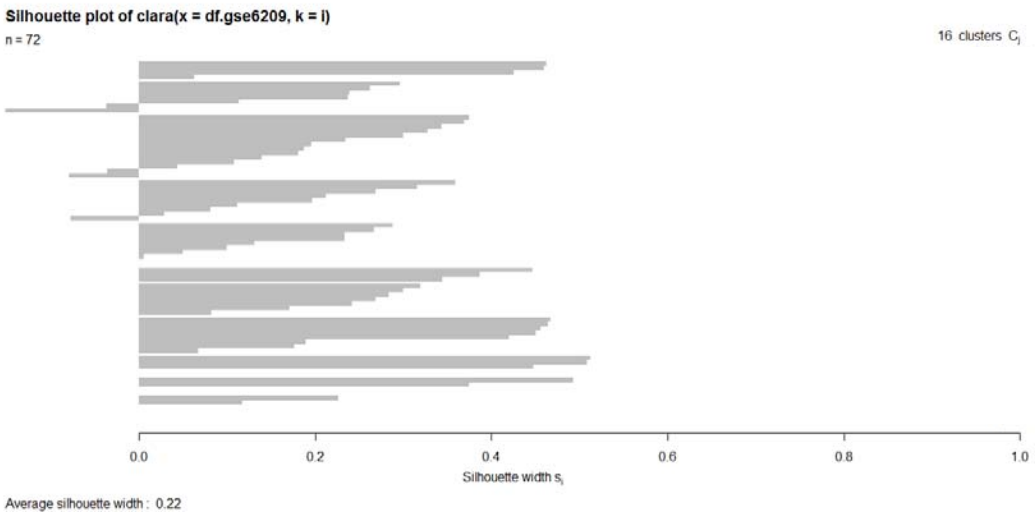


Figura 23. Silhoutte para k=16

2.3.2.2. Agrupamientos filtrados con ANOVA utilizando el algoritmo HOPACH

De la misma manera que para el caso anterior, se aplican los filtros a los experimentos, pero en este caso se utiliza el algoritmo HOPACH para el análisis de agrupamiento. En la figura 24 se muestra el *silhoutte* para el experimento GSE6209, esta vez resultante de aplicar el algoritmo HOPACH al experimento filtrado con anova. Se puede observar que si bien mejora con respecto al caso sin filtro anova, la calidad sigue siendo inferior al

agrupamiento conseguido mediante CLARA. Se pueden ver varios grupos con muchos de sus elementos con *silhoutte* negativo, indicando que deberían pertenecer a algún otro grupo en lugar del seleccionado por el algoritmo y un *silhoutte width* promedio bajo. Sin embargo hay aún grupos pequeños con un *silhoutte* razonable (alrededor de 0,3), adecuados para ser tenidos en cuenta en un análisis posterior.

El detalle con el análisis para el resto de los experimentos se muestra en el apéndice D.

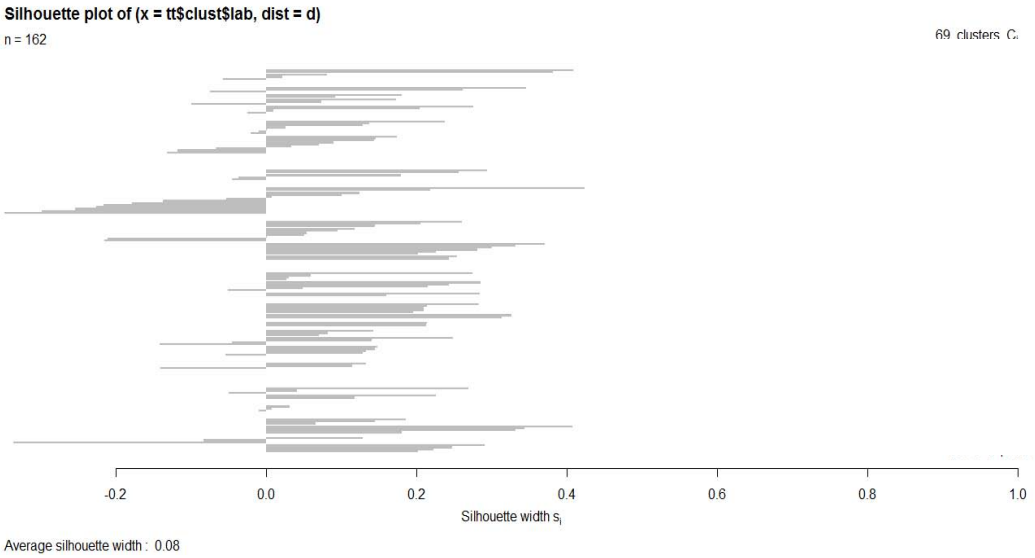


Figura 24. Silhoutte (69 grupos determinados por HOPACH)

Para finalizar con el tema de agrupamientos sobre experimentos cuyos datos fueron filtrados mediante anova, consideremos la tabla siguiente, que resume los resultados obtenidos, tanto utilizando el algoritmo CLARA como el HOPACH.

Agrupamiento convencionales con filtro anova						
Identificador del Experimento	**# Muestras**	**Método**	**# Genes que pasan el filtro**	**# Grupos**	**Tamaño promedio de los grupos**	**Tamaño del grupo mayor**
GSE6209	**11**	CLARA	162	16	27	55
		HOPACH	162	69	7.66	11
GSE12364	**12**	CLARA	62	11	15.09	82
		HOPACH	62	26	3.71	14
GSE15976	**36**	CLARA	2611	120	21.7	224
		HOPACH	2611	188	13.88	317
GSE9776	**17**	CALRA	273	30	9.1	42
		HOPACH	273	27	10.11	92
GSE365	**26**	CLARA	543	30	18.1	109
		HOPACH	543	19	28,57	195
GSE7962	**23**	CLARA	495	30	16.5	95
		HOPACH	495	38	15.46	150

Tabla 10. Agrupamientos convencionales con filtro Anova. Se muestra, para cada experimento, el número de muestras que lo componen, los métodos de agrupamiento convencionales utilizados, la cantidad de genes que se pasan el filtro Anova, la cantidad de grupos encontrados para el conjunto filtrado por Anova utilizando el algoritmo mencionado, el tamaño promedio de los grupos conseguidos y el tamaño del mayor de los grupos.

2.3.3. Selección de grupos de genes desde los agrupamientos convencionales

En este apartado se muestran algunos grupos de genes potencialmente útiles para encontrar patrones interesantes de acuerdo a sus características estadísticas. Los grupos que se buscan son aquellos para los cuales sus genes poseen el mismo patrón de expresión, infiriendo que pueden tener los mismos factores que lo controlan. De esta manera se convierten en buenos candidatos para explorar regiones promotoras y buscar patrones que sean zonas de reconocimiento y unión de proteínas que regulan la transcripción.

Del agrupamiento convencional con CLARA sin filtro anova para el experimento GSE6209 se selecciona un grupo con el valor de *silhouette width* máximo (0,955) y que además contiene un número de genes adecuado para la búsqueda de patrones (46 genes). Como se mencionó anteriormente, se consideran adecuados grupos de entre 20 y 80 genes. Para cada agrupamiento, además de seleccionar grupos con tamaños pequeños, de acuerdo al criterio anterior, se eligen aquellos con un *silhouette width* adecuado (*silhouette width* mayor a 0,2).

A este grupo de genes lo denominaremos GG 6209 Clr.1 (por el experimento, el algoritmo utilizado, y número de versión dado que más adelanta se utilizará el mismo experimento con el mismo algoritmo pero aplicando un filtro de *anova*), y está compuesto por los genes siguientes:

GG 6209 Clr.1
Rv0047c Rv0113 Rv0190 Rv0201c Rv0277c Rv0432 Rv0652 Rv0792c Rv0956 Rv1041c Rv1050 Rv1056 Rv1073 Rv1081c Rv1092c Rv1194c Rv1301 Rv1316c Rv1347c Rv1543 Rv1704c Rv1736c Rv1750c Rv1765c RV1828 Rv2034 Rv2042c Rv2170 Rv2215 Rv2267c Rv2282c Rv2377c Rv2424c Rv2560 Rv2596 Rv2703 Rv2856 Rv2954c Rv3217c Rv3234c Rv3310 Rv3514 Rv3533c Rv3737 Rv3874 Rv3891c

Del agrupamiento convencional con CLARA sin filtro *anova* para el experimento GSE10391 se selecciona un agrupamiento con un *silhouette width* de 0,7618 y conteniendo 15 elementos.

GG 10391 Clr.1
Rv0176 Rv2979c Rv3606c Rv1334 Rv0292 Rv3799c Rv0016c Rv1338 Rv2328 Rv1842c Rv1677 Rv2860c Rv0500 Rv1605 Rv0248c

Del agrupamiento convencional con HOPACH sin filtro a*nova* se puede seleccionar un grupo con un *silhoutte width* de 0,31 (con HOPACH esta

medida resultó ser siempre inferior comparada con CLARA), que consta 6 genes rotulados como GG 6209 Hpc.1, si bien este tamaño podría resultar pequeño para el análisis de patrones.

GG 6209 Hpc.1
Rv0932c Rv1045 Rv1509 Rv1902c Rv2171 Rv2882C

Del agrupamiento convencional con HOPACH sin filtro *Anova* para el experimento GSE12364 se puede seleccionar un grupo de 11 genes con un silhoutte witdh de 0,22.

GG 12364 Hpc.1
Rv1285 Rv1806 Rv1807 Rv1813c Rv2007c Rv2029c Rv2031c Rv2323c Rv2623 Rv2626c Rv2662 Rv3130c Rv3189

Del agrupamiento convencional con CLARA con filtro *anova* para el experimento GSE6209 se selecciona el grupo 12, con un *silhoutte width* de 0.43 y compuesto de 6 genes.

GG 6209 Clr.2
Rv1984c Rv2704 Rv3193c Rv0769 Rv1945 Rv1403c

Del agrupamiento convencional con HOPACH con filtro *anova* para el experimento GSE6209 se selecciona un grupo con un *silhoutte width* de 0.28 y compuesto por 6 genes.

GG 6209 Clr.3
Rv3606c Rv0568 Rv0513 Rv1257c Rv2217 Rv1333

Finalmente, del agrupamiento convencional con CLARA para el experimento GSE15976 se selecciona el grupo siguiente.

GG 15976 Clr.1
Rv0847 Rv2745c Rv1285 Rv0351 Rv2694c Rv2744c Rv3334 Rv2050 Rv1286 Rv1460 Rv0350 Rv0352

2.3.4. Biagrupamiento

En esta sección se muestran los resultados de agrupar los experimentos mediante biagrupamiento. Como se mencionó anteriormente se utilizaron los métodos CC, Plaid, Bimax y Quest, descriptos en la sección 2.5.

Aunque en cada caso se aplicaron todos los algoritmos, se presentan solamente los resultados de los métodos que dieron algún resultado relevante, junto a una tabla donde se pueden apreciar los valores de las medidas presentadas en el apartado teórico, y finalmente gráficos, de calor y de coordenadas paralelas, donde se resaltan las características principales del agrupamiento. En todos los casos se utilizaron los experimentos con el filtro *Anova* aplicado. En el caso de los mapas de calor, el primero de todos grafica la totalidad de los genes para todas las condiciones, mostrando la ubicación de los todos los biagrupamientos encontrados.

Utilizando el método Bimax para el experimento GSE6209, se obtienen los resultados mostrados en la tabla siguiente. Por cada biagrupamiento encontrado figura el número de filas y número de columnas que lo definen, y las validaciones correspondientes: valores constantes, coherente aditivo y coherente multiplicativo.

Bicluster Bimax para el experimento GSE6209					
Identificador	# Fil.	# Col.	Validación		
			Const.	Adit.	Mult.
BC1	**5**	6	1.38	0,70	9,54
BC2	**3**	7	1.42	0,76	0,43
BC3	**5**	6	1.51	0,97	0,72
BC4	**5**	6	0.94	1,03	0,71
BC5	**4**	7	0.91	0,95	0,66

Tabla 11. Bigrupos encontrados por el algoritmo Bimax aplicado al experimento GSE6209. Se observan el identificador del bigrupo, el número de filas, el número de columnas, y las validaciones para los distintos tipos de biagrupamientos: constante, aditivo y multiplicativo.

Sabiendo que, un valor por encima a 1-1,5 es suficiente para determinar que el biagrupamiento no es constante ni coherente (elegimos por lo tanto valores por debajo de 1 para considerarlos constantes o coherentes), se puede apreciar que BC1 y BC2 pueden ser coherente aditivos, que BC2, BC3, BC4 y BC5 pueden ser coherentes multiplicativos, y que BC4 y BC5 pueden ser valores constantes (tanto en filas como en columnas). Este último hecho puede ser verificado en los mapas de calor correspondientes. Estos mapas de calor tienen los genes que componen cada uno de los biagrupamientos como filas y los nombres de las muestras como columnas. En la figura 25 se observa el mapa de calor correspondiente a todo el microarreglo en primer lugar, y continuando el mapa de calor de cada uno de los biagrupamientos numerados del 1 al 5. Se puede observar que mientras que BC4 y BC5 muestran valores constantes (tienen colores similares), no sucede lo mismo para el resto de los biagrupamientos. En cambio, para el caso de BC2, las características del mapa de calor del biagrupamiento, solo se puede explicar por un efecto aditivo o multiplicativo (se ven valores constantes por filas representados en el mapa de calor por zonas de igual color).

El mismo efecto se puede apreciar en un gráfico de coordenadas paralelas. En esta representación los genes corresponden a las variables y cada línea a los distintos experimentos, en donde las líneas negras representan los elementos constituyentes del biagrupamiento, en contraposición a las grises que representan elementos que no forman parte del biagrupamiento. En la figura 26 se muestran las gráficos de coordenadas paralelas para cada uno de los biagrupamientos, numerados de la misma manera que mencionamos anteriormente. Se puede observar que los gráficos correspondientes a los biagrupamientos BC4 y BC5 diferencian mejor la regularidad constante de las expresiones génicas de las muestras dentro del biagrupamiento (línea de color negro) de las muestras que se encuentran fuera del biagrupamiento (línea de color gris).

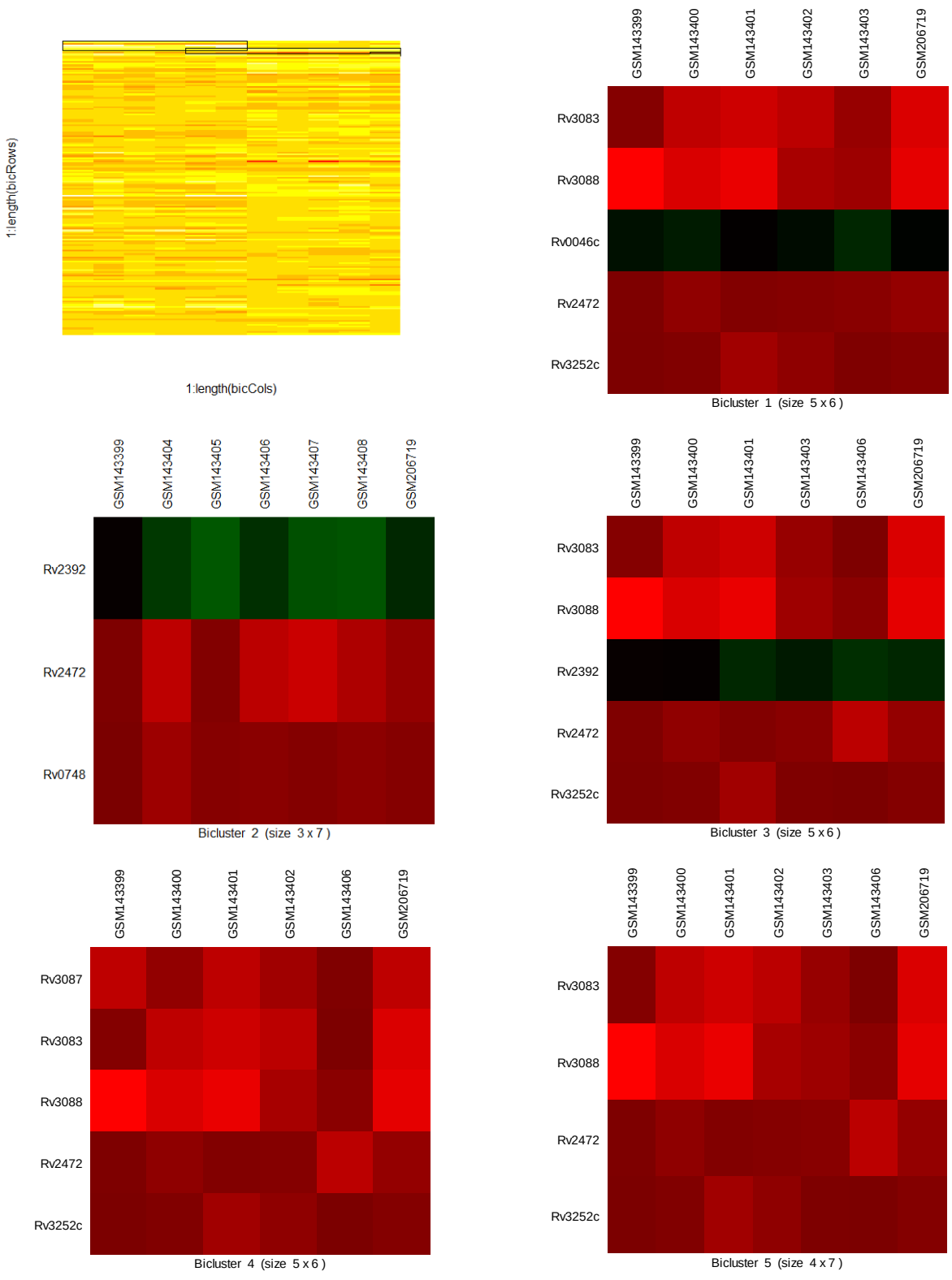


Figura 25. Mapas de calor para los biagrupamientos por el método Bimax

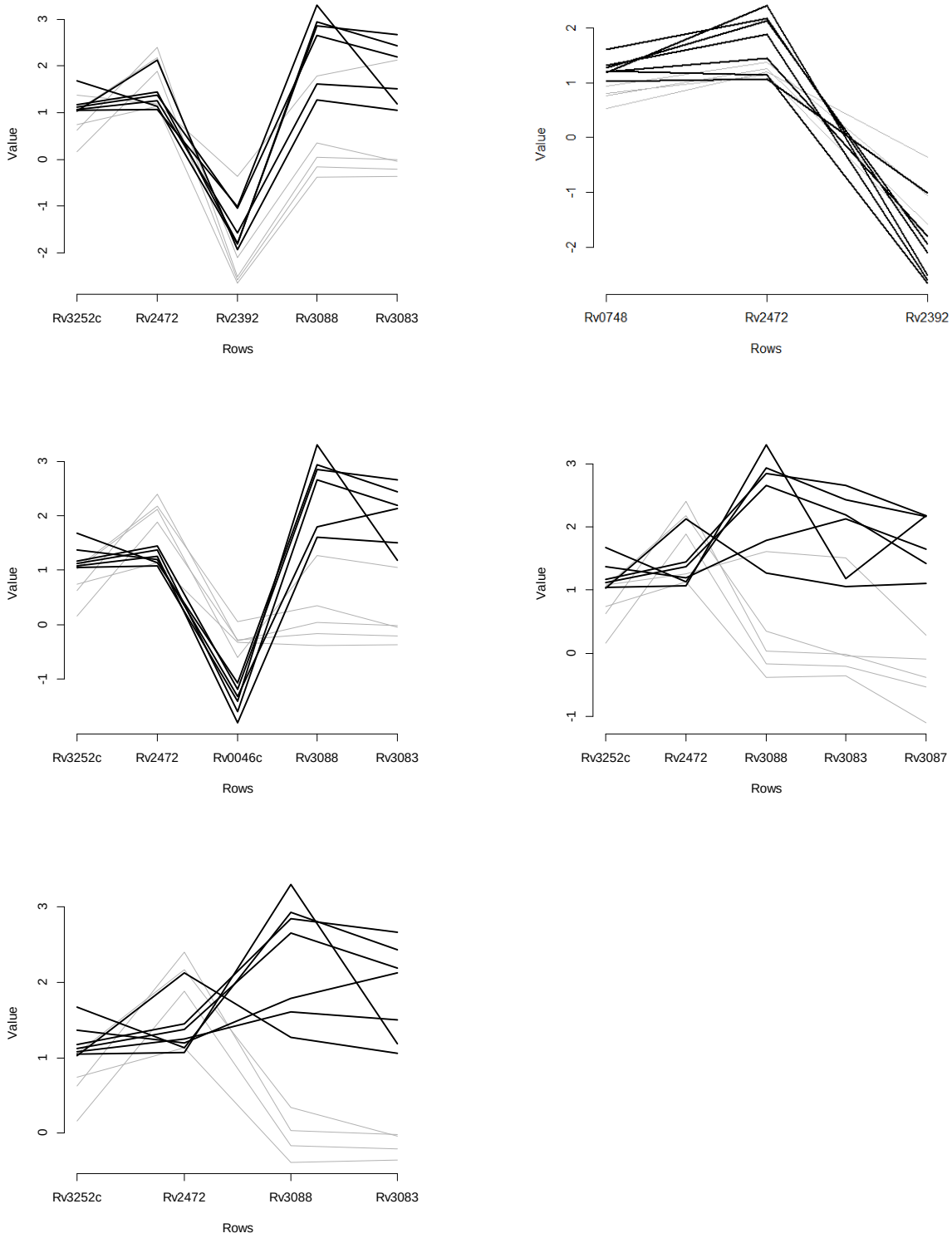


Figura 26. Coordenadas paralelas para los biagrupamientos por el método Bimax

Utilizando el método CC, este encuentra solamente un biagrupamiento que involucra a todos los genes y todas las condiciones, motivo por el cual no presentamos ningún resultado para el mismo.

Utilizando el método Plaid, se obtienen resultados con la condición de poner el parámetro background en FALSE. Si este parámetro TRUE, indica que existe un background constante para todas las filas y columnas. Este parámetro indica si el fondo (background), que es constante para todas las filas y columnas, está presente (TRUE) o no (FALSE) en la matriz de datos. A continuación se muestra la tabla conteniendo los resultados de aplicar este método.

Bicluster Plaid para el experimento GSE6209					
Identificador	# Fil.	# Col.	Validación		
			Const.	Adit.	Mult.
BC1	**25**	5	0,62	0,59	3,67
BC2	**15**	6	1,21	1,06	1,66
BC3	**11**	8	0,78	0,60	8,69
BC4	**26**	7	0,57	0,59	1,55

Tabla 12. Bigrupos encontrados por el algoritmo Plaid aplicado al experimento GSE6209. Se observan el identificador del bigrupo, el número de filas, el número de columnas, y las validaciones para los distintos tipos de biagrupamientos: constante, aditivo y multiplicativo.

El experimento GSE6209 tiene 11 muestras, divididas en dos grupos de 6 y 5 réplicas cada uno (ver la tabla 13, donde a fines explicativos se reitera la información de la tabla 7). El primer grupo corresponde a las condiciones “4 hs. después de la infección” y el segundo a “24 hs. después de la infección”. Sería trivial encontrar biagrupamientos donde las condiciones correspondan solamente a un grupo de condiciones (como es el caso de BC1 que abarca las 5 réplicas de la condición “24 hs. después de la infección”). Por otro lado uno esperaría en un caso ideal, que si un biagrupamiento tiene algunos casos de

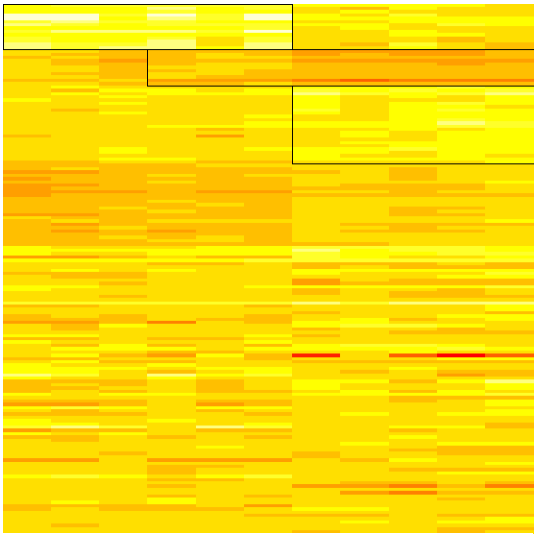
una condición, tenga la totalidad de casos para esa condición. Pero este caso ideal no siempre se cumple.

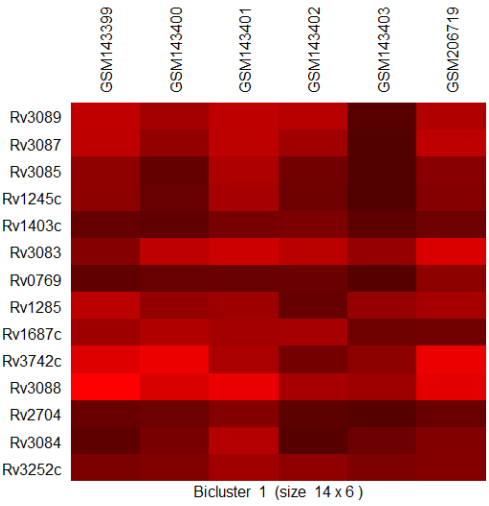
Si tomamos el tercer biagrupamiento, este abarca ocho columnas, conteniendo representantes de ambas condiciones, lo cual lo hace interesante para seguir analizando. Este grupo posee además una cantidad razonable de genes (11), y buenos indicadores de varianza constante y varianza aditiva (recordar que para que esto se cumpla los valores deben ser menores a 1 - 1.5, y se puede ver en la columna de validación tanto constante como aditiva que se cumple para el grupo mencionado).

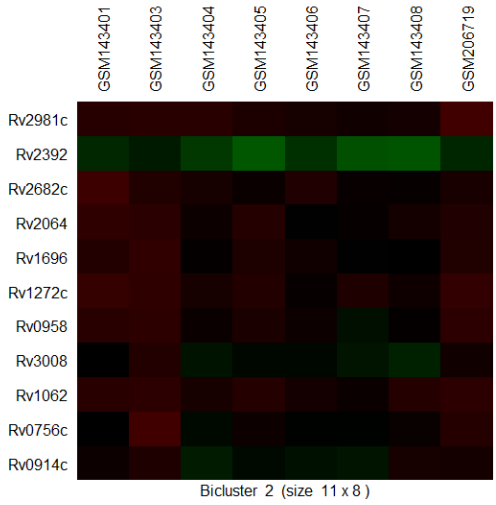
Muestras del experimento GSE6209	
Muestra	Descripción
GSM143399	**H37Rv 4 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 1 (A 4 horas)**
GSM143400	**H37Rv 4 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 2 (B 4 horas)**
GSM143401	**H37Rv 4 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 3 (C 4 horas)**
GSM143402	**H37Rv 4 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 4 (D 4 horas)**
GSM143403	**H37Rv 4 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 5 (E 4 horas)**
GSM143404	**H37Rv 24 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 1 (A 24 horas)**
GSM143405	**H37Rv 24 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 2 (B 24 horas)**
GSM143406	**H37Rv 24 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 3 (C 24 horas)**
GSM143407	**H37Rv 24 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 4 (D 24 horas)**
GSM143408	**H37Rv 24 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 5 (E 24 horas)**
GSM206719	**H37Rv 4 horas después de la infección de los macrófagos contra H37Rv cultivado en medio 7H9 réplica biológica 6 (F 4 horas)**

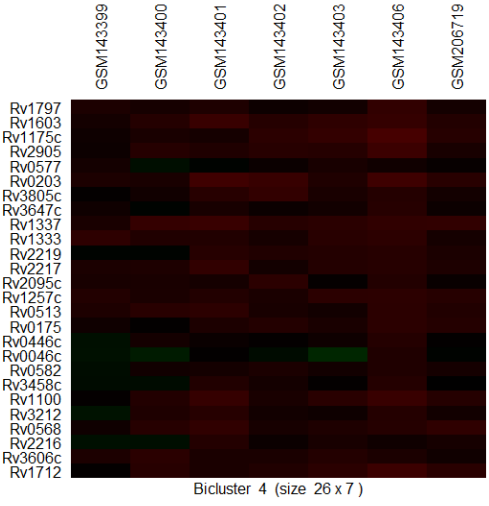
Tabla 13. Las 11 muestras del experimento GSE6209 con sus respectivas descripciones.

Por último en la figura 27 se muestran los mapas de calor y gráficos de coordenadas paralelas correspondientes a los biagrupamientos encontrados.









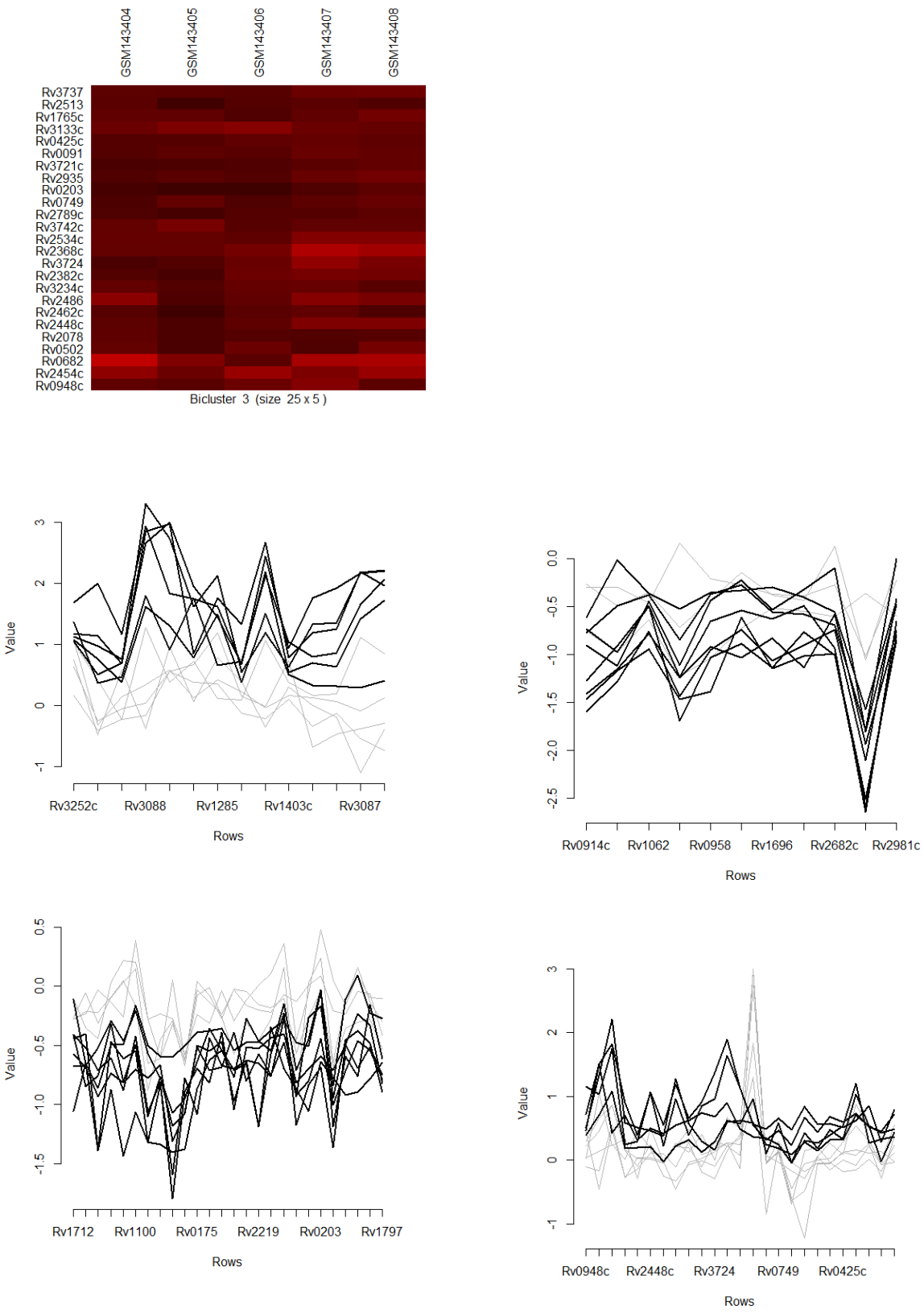


Figura 27. Mapas de calor y coordenadas paralelas para los biagrupamientos por el método Plaid

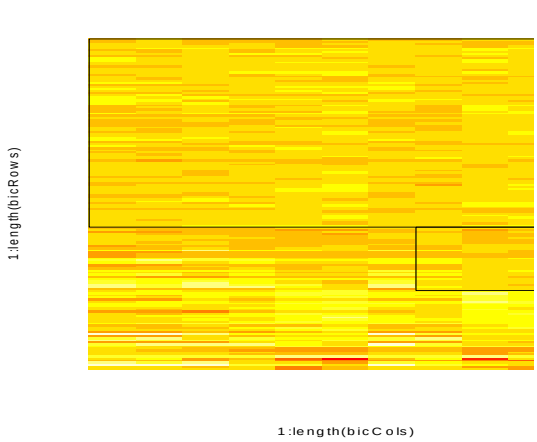
Por último, observemos los biagrupamientos por el método Questord, donde para nuestros fines, el único razonable para separar es el último grupo con 15 genes y todas las condiciones, dado que los demás tienen una cantidad muy grande de elementos, y el segundo además una cantidad muy pequeña de columnas. Además el último tiene medidas de valores constantes y coherencia aditiva por debajo de uno (condición que puede apreciarse en el mapa de calor de la figura 28).

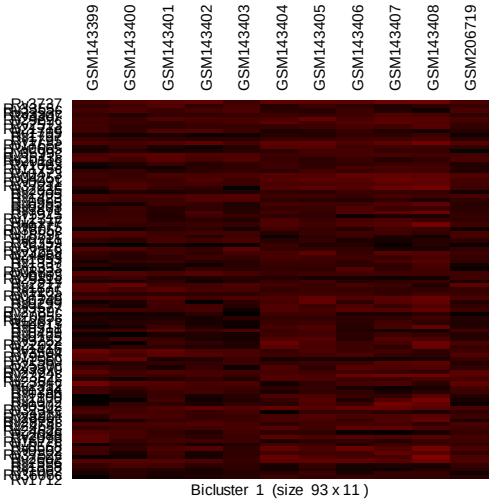
A continuación se muestra la tabla con los resultados de aplicar este método, y posteriormente en la Figura 28 los mapas de calor y gráficos de coordenadas paralelas.

Bicluster Questord para el experimento GSE6209					
Identificador	# Fil.	# Col.	Validación		
			Const.	Adit.	Mult.
BC1	**93**	11	0,99	1,00	Inf.
BC2	**38**	3	0,55	0,45	Inf.
BC3	**15**	11	0,94	0,90	13,75

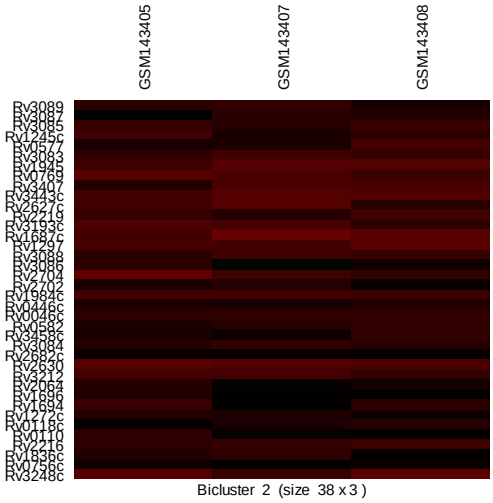
Tabla 14. Bigrupos encontrados por el algoritmo Questord aplicado al experimento GSE6209. Se observan el identificador del bigrupo, el número de filas, el número de columnas, y las validaciones para los distintos tipos de biagrupamientos: constante, aditivo y multiplicativo.

Como para las secciones anteriores, referirse a los apéndices para ver los resultados del resto de los experimentos.

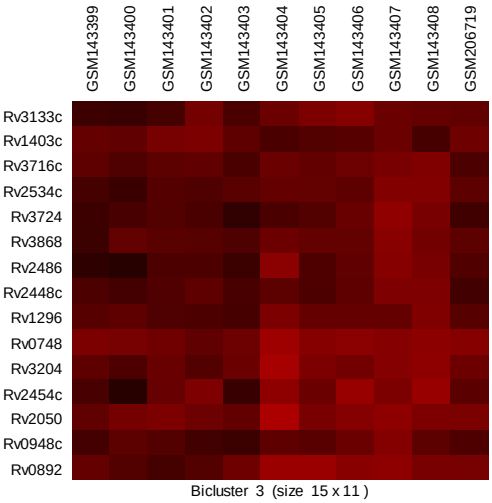




Bicluster 1 (size 93 x 11)



Bicluster 2 (size 38 x 3)



Bicluster 3 (size 15 x 11)

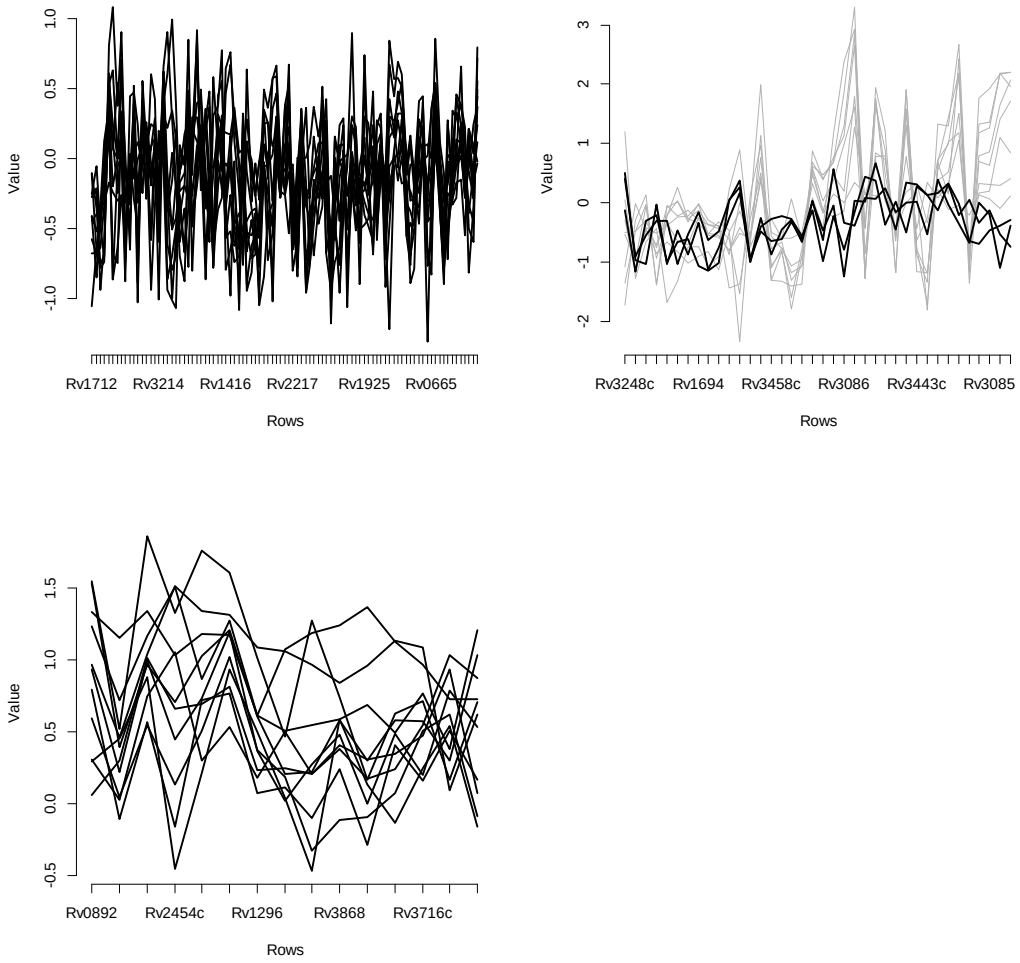


Figura 28. Mapas de calor y coordenadas paralelas para el biagrupamiento por el método Questord

2.3.5. Selección de grupos desde los biagrupamientos

Nuevamente, como en el caso de los agrupamientos convencionales, se espera que los genes pertenecientes a un mismo grupo sean genes con el mismo patrón de expresión, y por lo tanto pueden tener los mismos factores que lo controlen. Además recordemos que buscamos una cantidad pequeña de genes, para poder aplicar los métodos de búsqueda de patrones, y en este caso particular de biagrupamientos, que la cobertura sobre las condiciones

sea lo más amplia posible, lo cual podría indicar una mayor certeza en cuanto a que los genes comparten un mismo perfil de expresión.

Para el biagrupamiento del experimento GSE6209 por el método Bimax, se selecciona un grupo de 4 genes y una cobertura de 7 condiciones, siendo una buena cobertura sobre las 11 condiciones, pero una cantidad pequeña de genes.

GG 6209 Bimax.1
Rv3252c Rv2472 Rv3088 Rv3083

Para el biagrupamiento del experimento GSE6209 por el método Plaid se selecciona un grupo de 11 genes y una cobertura de 8 condiciones, resultando en una cantidad adecuada tanto de genes como de condiciones, con buenas medidas de valores constantes y coherencia aditiva.

GG 6209 Plaid.1
Rv0914c Rv0756c Rv1062 Rv3008 Rv0958 Rv1272c Rv1696 Rv2064 Rv2682c Rv2392 Rv2981c

Para el biagrupamiento del experimento GSE976 por el método Questmet se selecciona un grupo de 9 genes y una cobertura de 8 condiciones, dado que tiene grupos completos de columnas pertenecientes a diferentes condiciones, una cantidad razonable de genes, e indicadores de varianzas buenos para constantes y coherencia aditiva.

GG 976 quest.1
Rv0783c Rv1581c Rv0003 Rv0133 Rv0129c Rv3354 Rv1804c Rv0410c Rv0488

De este último agrupamiento también se selecciona un grupo de 22 genes con una cobertura de 6 columnas.

GG 976 quest.2
Rv3115 Rv1209 Rv3077 Rv0165c Rv0927c Rv2033c Rv2463 Rv2485c Rv3595c "Rv3637 Rv2647 Rv0258c Rv0282 Rv1414 Rv1196 Rv2014 Rv3614c Rv0346c Rv0678 Rv0748 Rv0874c Rv1692

Por último seleccionamos del experiento GSE365 el biagrupamiento con 11 genes y una cobertura de 22 columnas (de las 26 en total que posee).

GG 365 quest.1
Rv2227 Rv2231c Rv3689 Rv2279 Rv0988 Rv2607 Rv2720 Rv2797c Rv3587c Rv2189c Rv0751c

2.4. Validación semántica

En esta parte del proceso se valida semánticamente los conjuntos de genes conseguidos y validados estadísticamente en la etapa previa.

Se utilizó la medida de similitud semántica Term Overlap (TO), basada en alineamiento de grafos, que hace uso de la lista de anotaciones GO para cada gen de *Mycobacterium*, tomando la ontología GO como grafo de referencia. Es necesario por lo tanto contar con dichas anotaciones para cada gen. Esta tarea fue llevada a cabo dentro del contexto del presente trabajo utilizando la herramienta Blast2GO.

Para medir la significancia de la similitud semántica se procedió a realizar una prueba de permutación. En esta prueba se calculó el TO para 6000 pares de genes de *Mycobacterium* elegidos al azar. Posteriormente se crearon grupos de tamaño entre 2 y 80, muestreados mil veces desde la población de TO obtenidos anteriormente. Finalmente se creó una tabla de valores críticos para distintos valores de p (0.95, 0.975, 0.99 y 0.999).

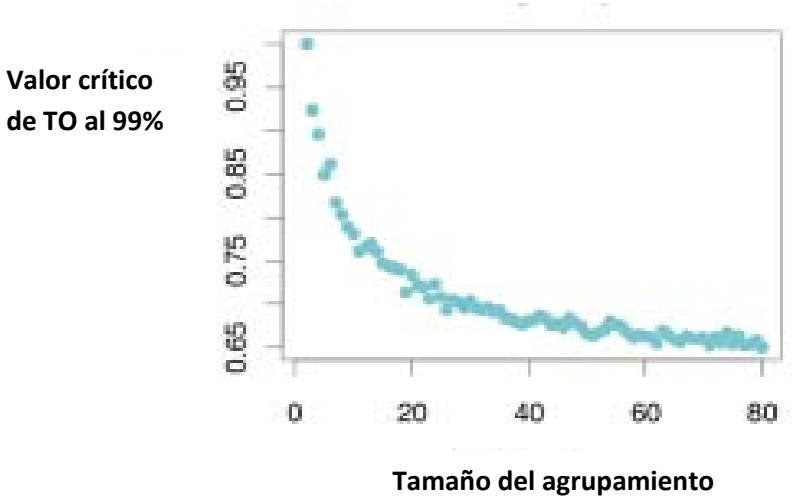


Figura 29. Valor crítico de TO al 99% para los distintos tamaños de agrupamientos. Se observa que a medida que el tamaño crece, el TO disminuye.

A continuación se muestra una tabla indicando la cantidad de grupos con TO significativos y no significativos, de acuerdo a cada una de las ontologías que componen GO (BP, CC y MF), para los agrupamientos conseguidos al aplicar el algoritmo CLARA a los experimentos filtrados por *anova*. Para cada grupo de cada agrupamiento se midió el TO del grupo de acuerdo a las 3 ontologías, y utilizando los valores críticos de la tabla construida anteriormente se determinó la significancia de cada grupo. El grupo se considera significativo si su TO es mayor que el valor crítico para el tamaño correspondiente.

Número de agrupamientos significativos (S) y no significativos (NS) para cada experimento									
Experimento	**BP**			**MF**			**CC**		
	NS	S	NA	NS	S	NA	NS	S	NA
GSE6209	8	4	0	11	1	0	8	1	3
GSE12364	1	4	6	1	4	6	2	2	7
GSE15976	40	36	44	50	10	60	30	50	40
GSE9776	2	2	26	2	2	26	2	3	26

GSE365	26	3	1	25	3	2	20	1	9
GSE7962	25	5	0	28	1	1	20	3	7

Tabla 15. Para cada experimento se muestra su identificador y la cantidad de grupos significativos y no significativos de acuerdo a su TO (grupos obtenidos al aplicar CALARA a los datos del experimento). NA significa que no se dispone de anotaciones para realizar el cálculo.

Además, en la tabla siguiente, se muestran el TO para los grupos de genes seleccionados anteriormente, seguido de su valor p de significancia. En cada caso se muestra el valor de TO calculado para el grupo seguido de su valor crítico. Los casos en donde el primer valor supera al segundo son significativos.

TO de los grupos seleccionados			
Grupo de Genes	**TO BP**	**TO MF**	**TO CC**
GG 6209 Clr.1	0,49 / 0,63	0,45 / 0,57	0,56 / 0,76
GG 10391 Clr.1	0,49 / 0,66	0,43 / 0,61	0,48 / 0,80
GG 6209 Hpc.1	0,40 / 0,73	0,62 / 0,65	0,46 / 0,85
GG 12364 Hpc.1	0,56 / 0,68	0,42 / 0,61	0,58 / 0,80
GG 6209 Clr.2	1 / 0,73	0,56 / 0.65	0,66 / 0,85
GG 6209 Clr.3	0,77 / 0,73	0,22 / 0.65	0,63 / 0,85
GG 15976 Clr.1	0,67 / 0,68	0,60 / 0,62	0,91 / 0,81
GG 6209 Bimax.1	1 / 0,76	0,66 / 0,68	0 / 0,87
GG 6209 Plaid.1	0,50 / 0,68	0,60 / 0,62	0,65 / 0,81
GG 976 quest.1	0,31 / 0,69	0,57 / 0,62	0,62 / 0,82
GG 976 quest.2	0,74 / 0,65	0,50 / 0,59	0,53 / 0,78
GG 365 quest.1	0,59 / 0,69	0,35 / 0,62	0,33 / 0,81

Tabla 16. Se muestra el identificador del grupo de genes seleccionados anteriormente, y para cada ontología el valor de TO seguido por su valor crítico (para el tamaño del grupo).

Finalmente, para aplicar el paso de búsqueda de patrones, se seleccionan algunos de los casos que son más significativos.

2.5. Búsqueda de patrones

En esta última etapa se procede a la búsqueda de patrones dentro de las regiones intergénicas correspondientes a los conjuntos de genes encontrados en el paso de agrupamiento y validados tanto numéricamente como semánticamente.

Previamente, utilizando las predicciones de DOOR y la ubicación de los genes (loci), se construye un archivo en formato FASTA conteniendo todas las regiones intergénicas. De esta manera, dado un grupo de genes obtenidos mediante los agrupamientos anteriores, obtenemos el conjunto de regiones intergénicas correspondientes.

Cabe esperar que los genes pertenecientes a un mismo grupo se expresen de manera similar, y por lo tanto sus regiones tengan patrones comunes que determinen este comportamiento. Se aplican por lo tanto algoritmos de búsqueda de patrones dentro de cada conjunto para validar esta hipótesis.

A continuación se muestran los resultados utilizando MEME para alguno de los conjuntos encontrados previamente. Este proceso se realiza con todos los grupos seleccionados, pero se muestran solamente los resultados más interesantes. Primeramente se muestra la secuencia de genes acompañada por su correspondiente región intergénica, y posteriormente los dos primeros motivos encontrados sobre estas regiones (expresión regular junto con un gráfico donde se visualiza dicha expresión y sitios dentro de las secuencias).

La tabla que sigue muestra para el grupo identificado como GG 15976 Clr.1, la secuencia *fasta* de las regiones intergénicas. Para cada gen perteneciente al grupo, esta secuencia *fasta* consta de el nombre del gen y debajo la secuencia intergénica correspondiente.

Rv0847

CGATCGCTCCTCGTCTGGATTTGGTCTCGCTCTTCGTACCCTGCCCAGACATCGGGCAGTACGCAACGGTTGATGATCACCACGCCATCATC
CGCCCTTACACCCTACCCCTATAGGGTATATAGTGGGCCACGTGGAAAGCGGGCAC

Rv2745c

TTGCGGGAACCAACCCCACCGCCGGCGGCGTTCACCTGATGACGATTGCGAGGTGGACAAGGAGTTTTTG

Rv1285

CAGCCCGAGCCGTCAGCCTAGGGCGCACTGGCGCACCGGCAGCCCGCCGAGATGGGGCTGCGTTGACAGCGATAGGGAAGCCTGGTTGCATA
G

Rv0351

GCACTGCTCTCCGGGCTTGACCGGGGCCTCTCCCAGCTACGCCCCGAGCGTGTGCCCTGCCGACACGCGGGAACAAGACCCGCACGACCAGC
GTTAGCATGCTCAGTAAGTTGAGTGCATCAGGCTCAGCTCTGAATTGACAGCACACCGCCGTCGAGGCAAGCTTGAGCGGGGTGCACTCATC
ATAGTGCAGGAAAGAAGCTCTACATATTCAGGAGGATTCACC

Rv2694c

CCCTTGATAGTGCCCGCCACTTTGGGGTGCTCCAAGGCCAACGGTGGCGCCCGGGAACCAACAGGTCTACGGTGGCCGTTGGCATAGGTGGA
ATACGCCGGCAGTTGCACGAAACTTCCATAACGTAGGTGACGTGTCAGGAGAGGCC

Rv2744c

GGGTCGCCGGTCGGTCCGGGTCGACAAGATGCGCGCCATGGGTGGGGCAATGCGCGCTGACCCGATAAATTGAGCGACAGGGCGCTGAGACC
ATCTGGCGCACACAAGAAGCGAAGGCGGAGCTAACTG

Rv3334

GATGGGGTTTCCGTTCTGCTAAAAGCCGTTACCTGGCGGGCTTTGGATCGCGATCCACGCCATAGGTGTGGCTGTCTGGTCAGGTTTGACCG
GCGCCATGATGTCGTTTCACAGCGCCGATGCAGTCTGGGAGGGGACCAGGGCATGGGTGCATTGAGGAGCCAGATCCAGAGAACCACACCGG
AGCCGCTGGCCGAGGCTCATCCACAAGCCTTCGATCCCGCTCCCGTTGTCGGCATGGGCGCCTGCCGACGGAATCAGCGGATGGTCATAGTG
GCGTCGGGCGCCAGGCCTGCGCGGGCACACGCGGTGCGGTGTCGATGGTTGTTCTCATCTGGTAACTCCTTTCCGCAGGCCGCAATTCAGCG
GTATGGGCTCACCGAGATCAGGCTCGTCACGATCGCCCGCACTGCTGGCGGCTCACATGTACCCAGTGTTAACCTTCTAGTGCACTAGAAGG
TCAAGGGGAGTCGC

Rv2050

GGCGCGCTGCTCGTAAGCCCGCTGCCGGCAAGACTGCCGGCAATACCGGCGGCGACGGCCCATGCCGACGTCGGTCACGTCACGGCCACACC
ACCCGCACGGCTGCGGACGGGCACGACGAGTCATGCCTGCAGACATTAGTCCGCCCGGGTGTCCGATCCCGGTATCATTGATGGTCGCGCCG
CGCGCGTCGCGTGCCGGGAACTACGCAGACGGCCGCAGCGTTTGCCAACCGGAGCCAGTCGCCAGTACGCAACCTACCAGCAGAGCCCAGGG
CTCACAGGACCTAAAGGAGTAGCGCCC

Rv1460

TGGTGCAGCCTAACGGCATGCCCGGGAATTGCTTAGGCGATCTCAAT

El primer motivo que aparece consta de 9 sitios, y la expresión regular encontrada es: A[TC][AG][ACG][CA][GT][TA][CAGT][GAT]G[CG]GA. Una expresión regular es una expresión que describe un conjunto de cadenas sin enumerar sus elementos, utilizando operadores especiales. Uno de estos operadores está compuesto por los símbolos [], indicando que a la cadena le sigue alguno de los caracteres mencionados. Por ejemplo, para cumplir con la expresión regular previa, la cadena debería comenzar con A, luego seguir con una T o una C, posteriormente seguir con una A o una G, y así siguiendo hasta completar la cadena. Otra forma de visualizar la expresión regular puede verse en la figura 30, en donde se muestra como una secuencia de logos. En este gráfico cada posición está compuesta por una pila de letras, donde la

altura de la pila es el contenido de información de esta posición en el motivo en bits y la altura de la letra individual en la pila es la probabilidad de la letra en esa posición multiplicada por el contenido de información total de la pila.

En la figura 31 se muestra una tabla conteniendo para cada gen dentro del grupo que comparten el motivo la siguiente información: el nombre del gen, el comienzo del motivo dentro del de la zona intergénica de ese gen, el valor p y la secuencia en donde este motivo aparece. En la figura 32 se muestra un diagrama de bloques, mostrando la localización del motivo dentro del gen como un pequeño bloque celeste, con su altura proporcional al -log(p-value).

Figura 30. Secuencia de LOGOS

Nombre	Comienzo	*p*-value	Sitios		
Rv2694c	120	7.53e-07	CGAAACTTCC	ATAACGTAGGTGA	CGTGTCAGGA
Rv2744c	65	1.95e-06	CGCTGACCCG	ATAAATTGAGCGA	CAGGGCGCTG
Rv2050	43	6.38e-06	ACTGCCGGCA	ATACCGGCGGCGA	CGGCCCATGC
Rv2745c	39	8.75e-06	CGTTCACCTG	ATGACGATTGCGA	GGTGGACAAG
Rv1460	28	1.45e-05	ATGCCCGGGA	ATTGCTTAGGCGA	TCTCAAT
Rv3334	120	2.48e-05	CACAGCGCCG	ATGCAGTCTGGGA	GGGGACCAGG
Rv0847	120	2.48e-05	CCTATAGGGT	ATATAGTGGGCCA	CGTGGAAAGC
Rv1285	66	8.51e-05	GGCTGCGTTG	ACAGCGATAGGGA	AGCCTGGTTG
Rv0351	145	1.26e-04	AATTGACAGC	ACACCGCCGTCGA	GGCAAGCTTG

Figura 31. Sitios encontrados

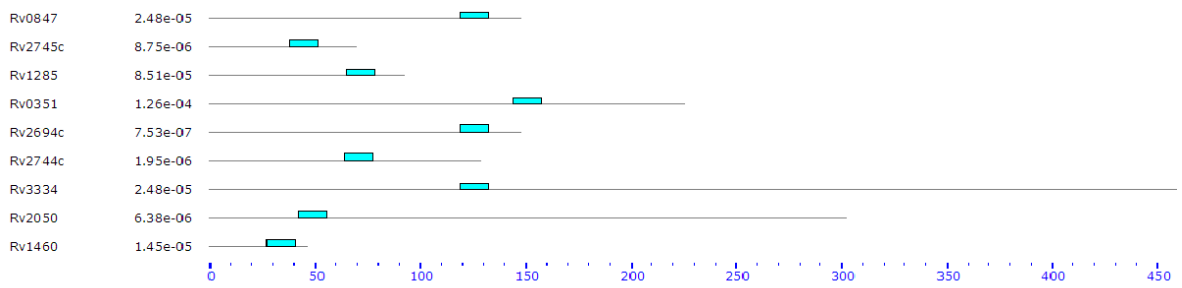


Figura 32. Diagrama de bloques

El segundo motivo está compuesto de 4 sitios, y su expresión regular es: CGGGAAC[CAT]A. En la figura 33 se puede observar la secuencia de logos, y en las figuras 34 y 35 se muestra la tabla con información para cada uno de los genes en donde el motivo aparece y el diagrama de bloques respectivamente.

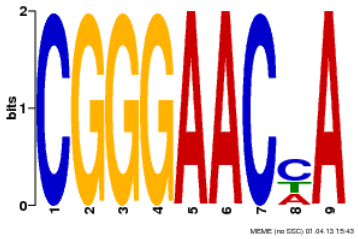


Figura 33. Secuencia de LOGOS

Nombre	Comienzo	*p*-value	Sitios		
Rv2694c	52	7.03e-06	CGGTGGCGCC	CGGGAACCA	ACAGGTCTAC
Rv2745c	4	7.03e-06	TTG	CGGGAACCA	ACCCCACCGC
Rv2050	199	1.11e-05	CGTCGCGTGC	CGGGAACTA	CGCAGACGGC
Rv0351	68	1.55e-05	TGCCGACACG	CGGGAACAA	GACCCGCACG

Figura 34. Sitios encontrados

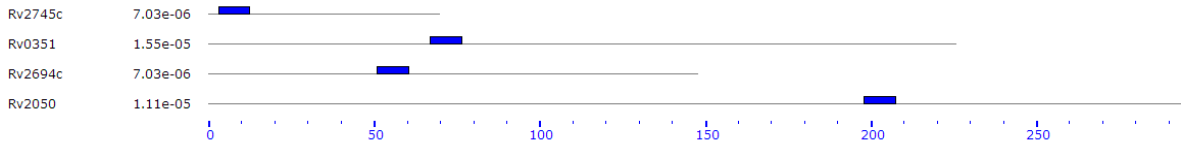


Figura 35. Diagrama de bloques

El mismo análisis se realiza para el grupo identificado como GG 6209 Plaid.1. Se muestra a continuación la secuencia fasta con las regiones intergénicas.

Rv0914c

AAGTCCGGCAACCGTTCGTCGCCGCGCGGAAAATGCCTGGTGAGCGTGGCTATCCGACGGGCCGTTCACACCGCTTGTAGTAGCGTACGGCT

Rv0756c

TGGTTGAACGTTACCTTCACAGTCATTGTGTAATTCCTGAAAGCTCGTTGCCAGTAGTCTGCTAACAGTCTGCCAGGAATCGCCAAATCAGC
TTGGACCGTTGCCGCTCAATCCACGGCGCGCCGTGAATACACTCGCAGA

Rv1062

CTCGAACGCGCTTCTCGGGGGAACCCGTTTCTCATGACTTCTCGCGGCGATAGCATTCGCCCGAGGAGGACATGAGGCGCGCCGAGACCCGT
AAGGCGGTACATCG

Rv3008

TCTTTCCGAGCGTTCAGTCGGCGAATCGCCGCTACCGCCATCACGTCGACCGGTGAGGCCGCCGTCACGGCCCCAAATCGGCGACGATCTGG
GCACGGAATCAGAACCTGATTGGGTCCCGGCCAGCCTCGCTGGCGTGGGAAGTCACCACGGTCGCGCCGCGGCTTCTCAACCCGGCCGACAC
GCGCTCCCG

Rv0958

GGCCGCTGGCCGCGACAGTGAAATCCACGACGTGACACGCCGGAAACGCGTCGTGACATTCACTCTCGTGGCCAGAAGAAAGACGGCGTCGT
AGCGTGGAACGGTG

Rv1272c

CGCTACCTCTCCCACTGCCGATCACAGCATGCACTGTAGCCGCAGGCCCGCGGGTCAACCACGACACGCCAGTTAACCAGGCATTTCGGCCT
AAGACTCATGCCTGGCGCGGGTAGGCAGCATGCGCAGTGTGCGACAGTTACCCA

Rv1696

TAGTTACGGGGCGCAGCAACCCCCGTAACCTCTACCGA

Rv2064

GCCACGGCGAGGAGACCTTTGGCTGGTCAGCCTCGGCGCCGCTCGCGCGGGTGAGCCCGGCAAGCATCGGCCCGCGGTGGTCGTTTCCGTGG
ACGAGCTACTCACCGGAATCGACGACGAACTCGTTGTCGTCGTGCCGGTGTCAAGCTCGCGCTCCCGCACCCCACTCCGGCCACCTGTCGCG
CCCTCAGAAGGTGTAGCTGCCGATAGCGTCGCGGTGTGCCGCGGCGTCCGCGCGGTCGCTCGTGCCCGACTCGTGGAGCGACTCGGCGCCCT
CAAACCCGCCACGATGCGCGCAATCGAAAACGCCCTGACCCTGATCCTCGGCCTCCCGACGGGACCTGAGCGCGGCGAGGCGGCGACCCATT
CTCCCGTACGGTGGACGG

Rv2682c

ATCGAACTCCGGAGTTGGATGACAATGAGTCGCCGACCAGGCTTCCCGGCGCGCCGGCCACCGACGCTATCAAAGTGGTTCGGCCCTTTTGA
TCGGCGGCCGCATGCTGGGCCAATGGGGGGGTTGCTGCGTACACTAGCGAA

Rv2392

TCTGCGCACGATTGACCTCGATCACTATCCGCTAAGACAACTATCTCAGTAGTCATATTTGGTCACATCTGTCACTCCTGTCAACGTCAGGT
GCGCGTCTCCCAGCGGATTCCCGGGTCGGCCTATCCATCCATCCAGGCTTGTTGCGTAGTTTTGATCATCGTGAAAAGAAATTTGACCAGGT
CGCGCAGCTGCACGCCATCCATGGCAGAATGTCACCGTGACCGCCGCCAAGAACCCGCGCCCCGATCTGCGAATCGCGCTGGTGGCTCGGCG
GCACATCGACCTCAAGCGGGTCTGCAGCTGTGGCTGTCGGCCTTGACGCCGTAAACCCAGCCCACCTGTATCTGCAGCCGGCGACCGGATCT
GCCCCTCCCGGAACAAGCGGCGTTTAGCGCGTCCTAGGTCGGCG

Rv2981c

GAAGTCATCCTGCCAGCGTCGATCCACGCGGCACACTTCGACGGCATTGCCGCCACGGTCGTGGCCGGGGCCCAGGCACGGTCCCGACGGCA
ACCGCGGCGCAGATTAGTGTGTGTGG

El primer motivo encontrado consta de 7 sitios y su expresión regular es: [AC]A[TA]C[CG][AG]CG[AG]C[GA]. En las figuras 36, 37 y 38 se muestran los gráficos correspondientes a este motivo.

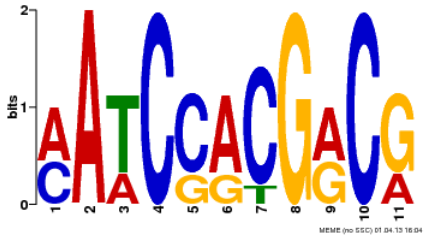


Figura 36. Secuencia de Logos

Nombre	Comienzo	*p*-value	Sitios
Rv0958	22	2.29e-07	GCGACAGTGA AATCCACGACG TGACACGCCG
Rv2064	109	9.53e-07	TACTCACCGG AATCGACGACG AACTCGTTGT
Rv0756c	110	9.53e-07	GTTGCCGCTC AATCCACGGCG CGCCGTGAAT
Rv3008	76	8.44e-06	CACGGCCCCA AATCGGCGACG ATCTGGGCAC
Rv1272c	56	8.70e-06	GCCCGCGGGT CAACCACGACA CGCCAGTTAA
Rv2392	200	2.20e-05	AGCTGCACGC CATCCATGGCA GAATGTCACC
Rv2981c	91	3.09e-05	GTCCCGACGG CAACCGCGGCG CAGATTAGTG

Figura 37. Sitios encontrados

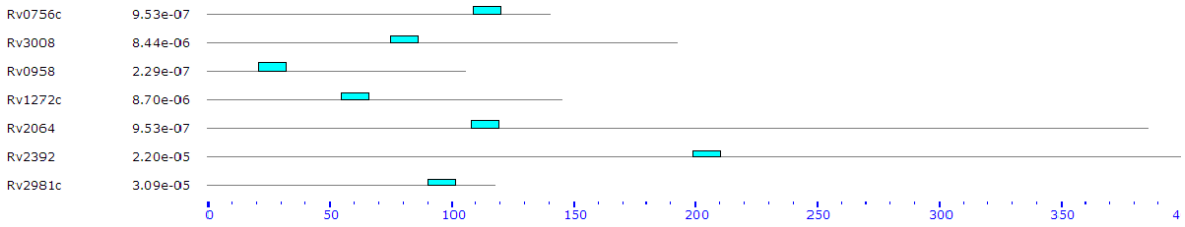


Figura 38. Diagrama de bloques

El segundo motivo encontrado consta de 11 sitios y su expresión regular es: [CG]G[CG]C[GA][TC][AGT][ACG]C[AGC][TG][CT]CAC. En las figuras 39, 40 y 41 se muestran los gráficos correspondientes a este motivo.

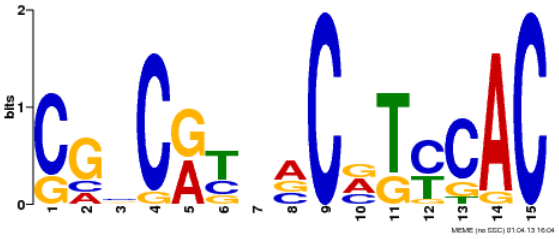


Figura 39. Secuencia de Logos

Nombre	Comienzo	*p*-value	Sitios
Rv0958	49	6.22e-08	GCCGGAAACG CGTCGTGACATTCAC TCTCGTGGCC
Rv0756c	6	3.95e-06	TGGTT GAACGTTACCTTCAC AGTCATTGTG
Rv3008	36	4.51e-06	TCGCCGCTAC CGCCATCACGTCGAC CGGTGAGGCC
Rv2981c	42	1.31e-05	CACACTTCGA CGGCATTGCCGCCAC GGTCGTGGCC
Rv1696	21	2.17e-05	GCGCAGCAAC CCCCGTAACCTCTAC CGA
Rv0914c	55	2.17e-05	CGTGGCTATC CGACGGGCCGTTCAC ACCGCTTGTA
Rv2392	272	3.43e-05	GCTGGTGGCT CGGCGGCACATCGAC CTCAAGCGGG
Rv2682c	47	8.79e-05	CCAGGCTTCC CGGCGCGCCGGCCAC CGACGCTATC
Rv1272c	20	1.08e-04	TCCCACTGCC GATCACAGCATGCAC TGTAGCCGCA
Rv2064	162	1.15e-04	CGCTCCCGCA CCCCACTCCGGCCAC CTGTCGCGCC
Rv1062	46	2.17e-04	GACTTCTCGC GGCGATAGCATTCGC CCGAGGAGGA

Figura 40. Sitios encontrados

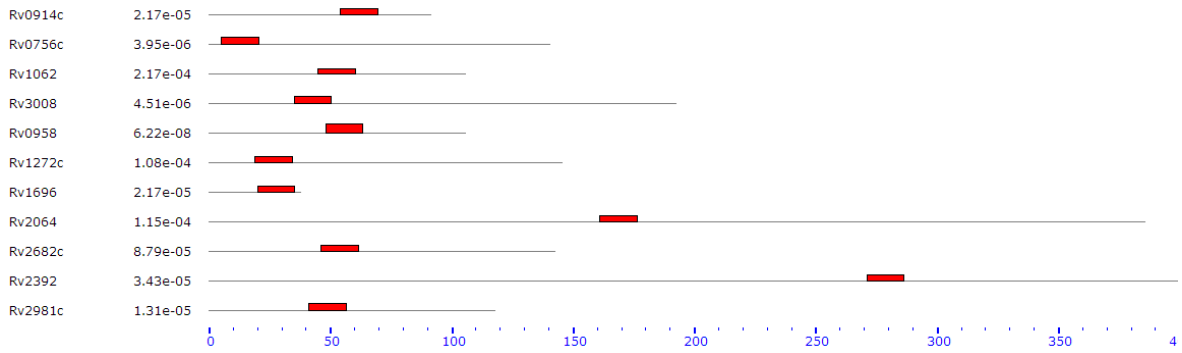


Figura 41. Diagrama de bloques

Finalmente, se realiza el mismo análisis para el grupo identificado como GG 976 quest.2. Se muestra a continuación la secuencia *fasta* con las regiones intergénicas.

Rv0165c

CTATCCCTTCCTTCCCTTCCTTTCACTTGCGCATCCCTTGCGCCAGGTTGATAT

Rv0258c

**CGCCACGCCCGAAGCCACAGAGGTGGGTATCGGCAATGGGCAATCCGGCAGCAATTGCCTGGTAACGCGACTGAAACCTCACAGGCCTAGAC
ACGTCAT**

Rv0282

**CGGCGCACCGTTCGCGCTGCCGGCACCCCGGGCTCCATAATGAAAATCATGTTCAGTAAGCTACACTCTGCATATCGGGCTACCAACGAAAT
GGAGTATCGGTCATGATCTTGCCAGCCGTGCCTAAAAGCTTGGCCGCAGGGCCGAGTCGATTGGTCGCGGTCGCCTCGACAGTTAGCTTATG
CAATGCTAACTTCGGGGCAAAGTTCAGGCGGATCGGCCG**

Rv0346c

**CGTCGCCTTCCGTCCGTTGATCTCACATCTCTGCGCTGGCGGATCAGGTGTCGGAGTCTCACACTACGACCGCGGCTACGACACCCTCGCGG
CCAGGACGGACCCGCACTTCGCGAGCGTCTGGCTGGCCAGGCGCGTCAGCCTGTGACCGCAATATCGACTGTCAACGCGGCCACCCGCGACG
GTCGGCACGACAACAAAGCACCGTTGACCGATCGATTGTTTCCGATTGGTTGCCAAGCGGCACCTCGCAGCCGCGTGGGATAGCCTGACAGC
CAGTCCGTTGGGTGTGCCCCATCTCACGCGCGTCCCAAGAAGCCACCATCCCAGGGAGCTCA**

Rv0678

CGCCCTCCGCCTCTGCCGCATGAAGTTCACGCCGGTCTGGTGACGCATACCGAACGTCACAGATTTCAGAGTACAGTGAAACTT

Rv0748

CCGGCGGGCCTGCGACCGCGCGCGGCGTTGACAGCATCGCTTCGGCCGCTCGACCGCAGATGATGCTGTTGATGCGTTACCGTGTGCATC

Rv0874c

**CGAGGCCGCGGCTCGACGGCCGACCGCGCTTAAGCGCGGTCGGCGCCAACGGTCCGAAGAGCCGCCGACACCCGGGGCACATCGGCGCATCA
TGGAACT**

Rv0927c

**TTCAGTTCCTAACGGCCGACCAGCTGTTTTGCTCGCAGCAAATGAAGCTAACAGCACTAGGTTAACGGTCGGCGAACTTGGCGTGCCGTCGC
CCTTGGCCGGGGTTGACCCCTGATGATGGACACGTGGCTGGTTCGACCGCGGGTCTCAGGGGAAGGATGTCCGTATGCCGTGCATACACAGC
AGGCACACCCCCAGGTTTCGGGAGCAGTCACAGGAAGCTGCAGCAAGCTTGTTACCGTTTCGGCCGACAGCCGGAAAAGAGGGCGCC**

Rv1196

GCCAGGAACAGTCGGCACGAGAAACCACGAGAAATAGGGACACGTA

Rv1209

CGGATTCCGGGCGCAGTCTCGATACCCGCACTGGACGCTCGACGGTAACCAGGCACTATGGATGC

Rv1414

**CATCGCTCGCCGAGGGAGGGAGCCCCATGTCTTGCATTCCGGACGAGATCGATACGCCCGACGTGCTGATCGACCGCGACATCCTTGACCGC
AACATCGGGCGAATGAGTTCCGCCGTCGCCGCGAAAGGGATCGCCCTGCGTCCCCACGTGAAGACGCACAAGCTGCCTGAGATCGCCCATAT
GCAACTCCGCGCGGGCGCGCGGCCTGACG**

Rv1692

TAGTTACGGGGCGCAGCAACCCCCGTAACCTCTACCGA

Rv2033c

**GAACCGTAAGTTTAGACTTACGGATATTCTTCCGCAAGGGTGCGCGGGTGGTGTGACACGTGATCCGGGCATCCCGTCACCCATGGCGTTGG
AGCCGAGTATCGGTTGGCGACGTTGGCCCGCCGGAAAGTCGACTGGTACAAGCATTTTCCGAGAACCCATGCGGCGCGTCGCGGCGATGACC
GGCAGAGAGGCTCCCTAGTATTGCGGC**

Rv2463

CGCTGACCTCGCGATCCTACGGCCCGGCGACACCGGCACCGCAATGACCTCTTGAGACCTCACGGGAAGGTCTCAAAACGACTCCGATTAGA
TTTGATGTCTGTCACCACGTACAGTCGCGCTCGACTAAATACAT

Rv2485c

GATGCGCAGGTCAGTTTATCTAACGGTGGCCGTCGAGGAATTTTTCCAGACGCCACACGCGCGAAAGGTCTACCGCCACGTCAATTATTGGT
TATTGTGTGATCCTCGACACACGTACAAAACCGGCTGGCAACCAGCCGTTTACCGCAACGATGCGTGTGCCTGACGTCGAGGGGTGGGCTTT
GTT

Rv2647

GTCGCCCTGGCGTTTGCTGCAGCCGATCGGCGTCACCCCCGACAGGCGGCTCGTATTCGGCCAGCGGCGGCTCGAGGCTGCACGGCTGCTCG
GATGGGAGCGCATCCCG

Rv3077

CTAGTGTCTCTTCGCTCAGTCGATGTCGACATTGTTCTCCTTAAACCGTAGCGACGTCGCAAATCGGATTGGCAGGATGCCCCGCAAAACCC
ACGTCC

Rv3115

CTCGGCTTGCCGACTACCTCACTGACCCAGGAGGAGAGTTACGTCCAGGGGTGTGGTGTACGGGCAGGTAAGGCCGGTGGGCGTGTCGTAGC
CCAGTAGTGGGCGGTCATCGCGTGATCCTTCGAAACGACCAGCAAAAGTCAATCGAAGGAAATGACGCA

Rv3595c

CGGTCTAGGTCCATCGGCCGTATTAGGCGAAAACCGCGACCGCCGATCCGCTGTGACAATAGACTCTGCTGACGTCGGTTTGTCCGGAAGCT
GTTCGGAGGTGGCCA

Rv3614c

CCACGTTTGCTGCCCGCAGTGCAGGCCACAGCGTCTTCCCAACGACCTGTTCGGACTGACCACGCCAGCTGCCCAGGCCGACCCTTCCCGGG
TGGCAATGAATTCCGAAGGGACG

Rv3637

GATCGGGCAGTGCGACTTCTGGTTCCCCGATGTCGTTGTGCCGGTGGGGTACGGCCAGGTCCGCACCGCCACGGCGTTACCTGTGCTGACCA
TGGTGTGTGGGTATTCGCGGTGGGCCTCGGCGCTGTTGATCCCGACACGCACCGCCGAAGACTTGTATGCCGGGTGGTGGCAGCATCTTTCG
ACGTTGGGCGCCGTTCCAAGGGTGTTGGTGTGGGACGGCGAGGGCGCGGTCGGGCGGTGGTGGGCGCGCCAACCTGAACTGACTGCGGCATG
CCATGCCTTCCGCGGCACCCTGGCCGCCAAAGTGTGGATCTGTAAACCGGTGATCCCGAAGCCAAGGGGCTGGTCGAACGTTTCCACGACTA
CCTGGAGCGGGCGTTC

El primer motivo encontrado consta de 6 sitios y su expresión regular es: GTGGG[GT][TG]ACG[GT]C[CG]AGG. En las figuras 42 y 43 se muestran los gráficos correspondientes a este motivo.

Nombre	Comienzo	p-value	Sitio		
Rv3637	44	2.90e-09	CGTTGTGCCG	GTGGGGTACGGCCAGG	TCCGCACCGC
Rv3637	212	1.02e-07	AAGGGTGTTG	GTGTGGGACGGCGAGG	GCGCGGTCGG
Rv3115	34	1.11e-07	GACCCAGGAG	GAGAGTTACGTCCAGG	GGTGTGGTGT
Rv3115	53	2.22e-07	GTCCAGGGGT	GTGGTGTACGGGCAGG	TAAGGCCGGT
Rv2485c	159	5.06e-07	AACGATGCGT	GTGCCTGACGTCGAGG	GGTGGGCTTT
Rv0258c	23	7.67e-07	AGCCACAGAG	GTGGGTATCGGCAATG	GGCAATCCGG

Figura 42. Sitios encontrados

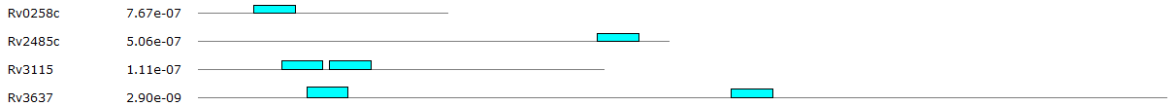


Figura 43. Diagrama de bloques

El segundo motivo encontrado consta de 7 sitios y su expresión regular es : ACCGC[AC]A[CT][GA][TA]C. En las figuras 44 y 45 se muestran los gráficos correspondientes a este motivo.

Nombre	Comienzo	p-value	Sitio
Rv1414	88	2.36e-07	GACATCCTTG ACCGCAACATC GGGCGAATGA
Rv0346c	148	7.55e-07	TCAGCCTGTG ACCGCAATATC GACTGTCAAC
Rv2463	38	2.36e-06	CGACACCGGC ACCGCAATGAC CTCTTGAGAC
Rv2485c	72	3.11e-06	CGAAAGGTCT ACCGCCACGTC AATTATTGGT
Rv1414	73	5.17e-06	GTGCTGATCG ACCGCGACATC CTTGACCGCA
Rv2485c	144	7.49e-06	CCAGCCGTTT ACCGCAACGAT GCGTGTGCCT
Rv3637	65	1.43e-05	CCAGGTCCGC ACCGCCACGGC GTTACCTGTG

Figura 44. Sitios encontrados

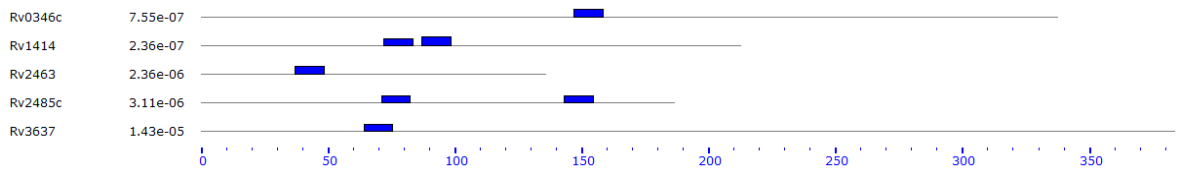


Figura 45. Diagrama de bloques

En este último caso se hizo también una búsqueda mediante MAST sobre las todas las regiones intergénicas de M. tuberculosis H37Rv, encontrando los motivos en algunos sitios de zonas intergénica de otros genes no pertenecientes al grupo bajo análisis. No necesariamente estos genes deberían estar dentro del grupo bajo análisis, dado que, por ejemplo, puede que no estén regulados en la misma manera. En la figura 46 se muestra el diagrama de bloques en donde se aprecia cada uno de los motivos (entre los que se encuentran los dos presentados anteriormente) y su localización en las zonas intergénicas.

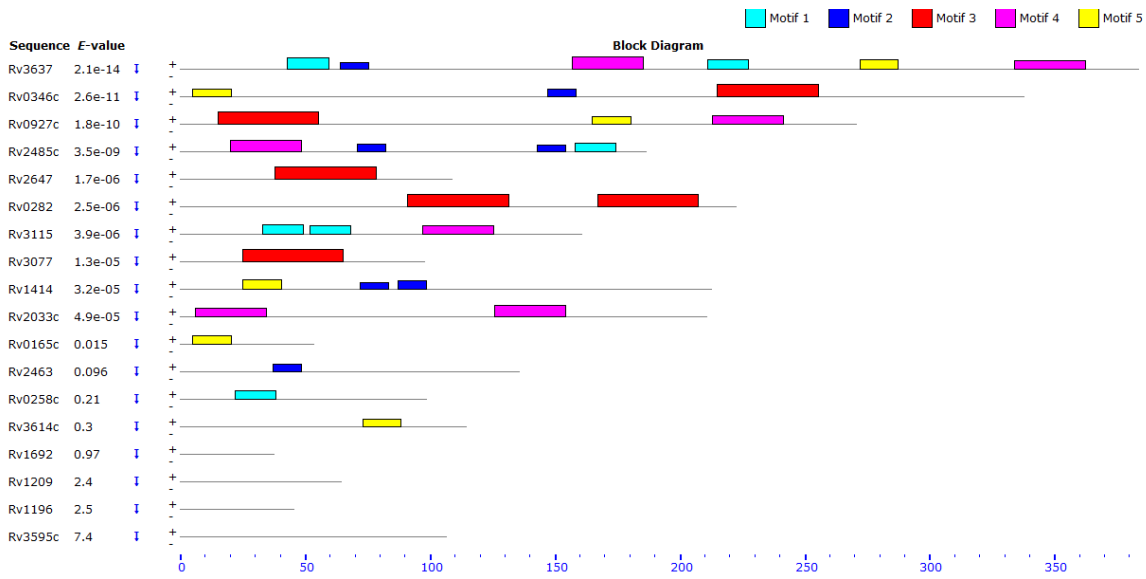


Figura 46. Diagrama de bloques

3. Conclusiones

Los experimentos con microarreglos son una poderosa herramienta para medir el nivel de expresión génica de los genes pertenecientes a un organismo simultáneamente. Estos niveles se pueden analizar por técnicas estadísticas diversas, pero una de las más utilizadas corresponden a métodos de agrupamientos, agrupando genes que tengan una misma respuesta.

Con el fin de conseguir estos agrupamientos se necesita realizar una serie de pasos, desde la obtención de los datos desde los repositorios correspondientes, el preprocesamiento, la ejecución de los algoritmos de agrupamiento con los parámetros adecuados y finalmente la validación tanto estadística como semántica de los grupos obtenidos.

Por otro lado, en el marco del presente trabajo, los grupos se utilizan para un análisis posterior: dada las regiones intergénicas de los genes pertenecientes a un grupo, determinar mediante algoritmos de búsqueda de patrones, motivos comunes que se pueden encontrar en las regiones promotoras. Cabe

esperar que los genes pertenecientes a un mismo grupo se expresen de manera similar, y por lo tanto sus regiones posean patrones comunes que determinen este comportamiento. Se aplican por lo tanto algoritmos de búsqueda de patrones dentro de cada conjunto para validar esta hipótesis.

Recordemos que el promotor de un gen es la región de ADN que controla la iniciación de la transcripción de dicho gen a ARN. Dicha región está compuesta por una secuencia específica de ADN localizado justo donde se encuentra el punto de inicio de la transcripción del ADN y contiene la información necesaria para activar o desactivar el gen que regula. En las regiones promotoras es común encontrar patrones cortos de secuencias de nucleótidos que se repiten en diferentes promotores y se denominan motivos.

3.1. El flujo de procesos

Una de las contribuciones de este trabajo es diseñar, evaluar y refinar un flujo de proceso, esquematizado en la figura 47, que involucra cada uno de los pasos mencionados anteriormente, para ser utilizado por los biólogos. Este flujo fue implementado principalmente en R, excepto la parte referida a la búsqueda de patrones.

En un primer paso se obtienen los datos desde el NCBI utilizando las funciones de R provistas en el paquete GEOQuery, y se transforman los objetos recuperados a objetos Expression Set, por ser estos la entrada de muchas de las funciones utilizadas posteriormente.

Como segundo paso se normalizan los datos, se analizan por casos nulos o extremos, y se aplican los filtros que se consideren pertinentes. En este trabajo se encontró que aplicar el filtro *anova* es adecuado en este punto del proceso. Debido a la gran cantidad de test involucrados, es posible encontrar algunos falsos positivos, pero el propósito del filtro es descargar datos con

ruido para la posterior aplicación de agrupamientos más que obtener una lista de genes diferencialmente expresados.

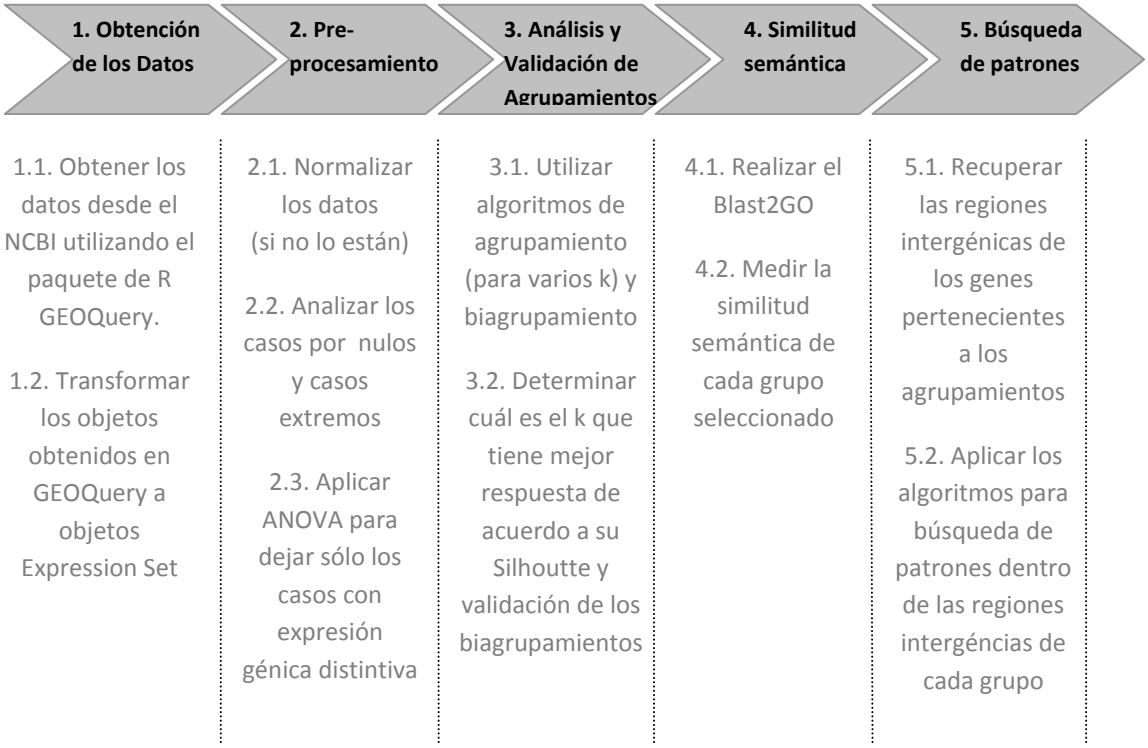


Figura 47. Flujo de procesos

El tercer paso es realizar los agrupamientos. En este trabajo se encontró que este paso puede realizarse de dos maneras efectivas: una es utilizar los algoritmos tradicionales CLARA y HOPACH, y otra es mediante los algoritmos más novedosos de biagrupamientos. Para el caso de CLARA, donde el algoritmo requiere el número k de grupos como entrada, se debe ejecutar el algoritmo para distintos k, comenzando por un valor cercano a 100 y terminando en un valor cercano a 200. De todos los agrupamientos, el elegido será el que tenga el mejor *silhoutte width*. Hay que tener en cuenta, sin embargo, que el *silhoutte* general puede no ser adecuado, y aún así tener algún grupo con un buen *silhoutte*, apto para ser seleccionado. Para el caso de biagrupamiento, se sabe que éste es más vulnerable a sobreajuste, y se debe por lo tanto garantizar que la solución hallada sea significativa

evaluando apropiadamente los resultados. Se recomienda por lo tanto utilizar las validaciones correspondientes a valores constantes, coherente aditivo y coherente multiplicativo, para determinar la validez del agrupamiento.

Como cuarto paso se realiza la similitud semántica basada en la medida de similitud *term overlap* sobre la ontología GO. Cabe aclarar que esta medida funcionaría mejor de haber una mayor cantidad de genes de *Mycobacterium tuberculosis* con anotaciones GO.

Como quinto paso se arman las secuencias *fasta* con las regiones intergénicas de los genes pertenecientes a los distintos grupos. Esta secuencia es provista al algoritmo de búsqueda de patrones MEME, con el objetivo final de encontrar motivos comunes dentro de las regiones promotoras de los genes que la componen.

3.2. Trabajos futuros

Hay algunos puntos más que podrían ser agregados al flujo de procesos en versiones futuras. Entre estos se pueden mencionar:

1. Aplicar algoritmos de consenso de agrupamientos. Estos algoritmos combinan el resultado de varios agrupamientos de manera de obtener un agrupamiento de mayor confiabilidad **[PON2011]** y **[KUN2004]**. Este tipo de procesamiento fue realizado en el transcurso de la tesis pero sin demasiado éxito, debido a que los algoritmos de consenso aplicados no mostraban buenos resultados.

2. Agregar técnicas de bootstrapping a los algoritmos de agrupamientos **[KAT2001]**, tanto convencionales como biagrupamientos. El bootstrapping es un método de remuestreo que permite resolver problemas relacionados con la estimación de intervalos de confianza o la prueba de significación estadística. Aplicado al análisis de agrupamientos, permite evaluar la estabilidad de los resultados.

3. Construir un conjunto de datos compuesto de varios experimentos y tomarlo como entrada del flujo de procesos. De esa manera también se consigue obtener resultados de mayor confiabilidad. Por ejemplo, se podrían unir las mediciones de cada uno de los experimentos utilizados como si se tratara de uno sólo. En este caso habría que determinar si se utilizaron los mismos criterios en la estandarización de los datos.

4. Bibliografía

[BAI1994] Fitting a mixture model by expectation maximization to discover motifs in biopolymers
Timothy L. Bailey y Charles Elkan
Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.
http://pages.bangor.ac.uk/~mas00a/papers/lkSMC04.pdf

[BAI2009] MEME SUITE: tools for motif discovery and searching
Timothy L. Bailey, Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li y William S. Noble
Nucleic Acids Res. 2009 Jul;37(Web Server issue):W202-8. doi: 10.1093/nar/gkp335. Epub 2009 May 20.

[COU2006] Measuring semantic similarity between Gene Ontology terms
Francisco M. Couto, Mário J. Silva, Pedro M. Coutinho
Data & Knowledge Engineering 61 (2007) 137–152

[DAM2007] Operon prediction using both genome-specific and general genome information
P. Dam, V Olman, K. Harris, Z. Su, Ying Xu
Nucleic Acids Research, 35:288 - 298, 2007

[DIV2011] An effective measure for assessing the quality of biclusters
Federico Divina, Beatriz Pontes , Raúl Giráldez y Jesús S. Aguilar-Ruiz
Computers in Biology and Medicine 42 (2012) 245–256
Noviembre 2011

[LI 2009] QUBIC: a qualitative biclustering algorithm for analyses of gene expression data
Guojun Li, Qin Ma, Haibao Tang, Andrew H. Paterson y Ying Xu
Nucleic Acids Research, 2009, Vol. 37, No. 15
Junio 2099

[FAL2006] An Introduction to Bioconductor's ExpressionSet Class
Seth Falcon, Martin Morgan, y Robert Gentleman
6 October, 2006; revised 9 February, 2007

[GEN2005] Bioinformatics and Computational Biology Solutions Using R and Bioconductor
Robert Gentleman, Rafael A. Irizarry, Vincent J. Carey, sandrine Dudoit y Wolfgang Huber
Springer - 2005

[GEN2009] Package genefilter
http://bioconductor.org/packages/2.3/bioc/html/genefilter.html

[SII2009] Visual Perception of Parallel Coordinate Visualizations
Harri Siirtola
Proceedings of the 2009 13th International Conference Information Visualisation, Pages:3-9

[HAS2011] Package impute
http://www.bioconductor.org/packages/release/bioc/manuals/impute/man/impute.pdf

[IVE2010] Un Acercamiento a la Ontología de Genes y sus Aplicaciones
Ivette Camayd Viera, Miguel Sautié Castellanos, María A. Zardón Navarro, Carlos Martínez Ortiz y C. José Luis Hernández Cáceres
Revista Cubana de Informática Médica Año 2010 Nro. 2

[JIA1997] Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy
Jay J. Jiang
Proceedings of International Conference Research on Computational Linguistics (ROCLING X), 1997, Taiwan.

[KAT2001] Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments
M. Kathleen Kerr y Gary A. Churchill
Communicated by Bradley Efron, Stanford University, Stanford, CA, May 30, 2001

[KAU1990] ***Finding Groups in Data: An Introduction to Cluster Analysis*****.**
Kaufman, L. y Rousseeuw, P.J.
Wiley, New York - 1990

[KEI2011] BiCluster Algorithms
Sebastian Kaiser, Rodrigo Santamaria, Tatsiana, Khamiakova, Martin Sill, Roberto Theron, Luis Quintales y Friedrich Leisch
http://cran.r-project.org/web/packages/biclust/index.html

[KUN2004] Using Diversity in Cluster Ensembles
Kuncheva, Ludmila I., Hadjitodorov, Stefan T.
IEEE SMC International Conference on Systems, Man and Cybernetics, 2004

[LIN1998] An information-theoretic definition of similarity
Lin, D. (1998)
In Proceedings of the 15th International Conference on Machine Learning, pp. 296–304.Morgan Kaufmann, San Francisco, CA. 1998

[MAD2004] Biclustering Algorithms for Biological Data Analysis: A Survey
Sara C. Madeira y Arlindo L. Oliveira
INESC-ID TECHNICAL REPORT 1/2004, JANUARY 2004 1

[MAH2004] Transcription factor binding site identification using the self-organizing map
Shaun Mahony, David Hendrix , Aaron Golden, Terry J. Smith y Daniel S. Rokhsar
Bioinformatics Volume 21, Issue 9 Pp. 1807-1814, 2004

[MAO2009] DOOR: a Database of prOkaryotic OpeRons.
F. Mao, P. Dam, J. Chou, V. Olman, Y. Xu
Nucl. Acids Res. 37: D459-D463, 2009

[MEE2008] Gene Ontology term overlap as a measure of gene functional similarity
Meeta Mistry and Paul Pavlidis
BMC Bioinformatics 2008, 9:327

[MUR2003] Extracting conserved gene expression motifs from gene expression data.
Murali TM, Kasif S.
Bioinformatics Program, 48 Cummington St., Boston University, Boston, MA 02152, USA.
Pac Symp Biocomput. 2003:77-88.

[PON2011] A SURVEY OF CLUSTERING ENSEMBLE ALGORITHMS
Sandro Vega-Pons y José Ruíz-Shulcloper
Int. J. Patt. Recogn. Artif. Intell. 25, 337 (2011). DOI: 10.1142/S0218001411008683

[PRE2006] A systematic comparison and evaluation of biclustering methods for gene expression data
Amela Prelic, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Bühlmann, Wilhelm Gruissem, Lars Hennig, Lothar Thiele y Eckart Zitzler
Bioinformatics Vol. 22 no. 9 2006, pages 1122–1129

[RES1999] Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language
Philip Resnik
Journal of Articial Intelligence Research 11 (1999) 95-130

[REY1992] Clustering rules: A comparison of partitioning and hierarchical clustering algorithms
Reynolds, A., Richards, G., de la Iglesia, B. y Rayward-Smith, V.
Journal of Mathematical Modelling and Algorithms 5, 475–504 - 1992

[STA2007] Biclustering in data mining
Stanislav Busygin, Oleg Prokopyev, Panos M. Pardalos
Computers & Operations Research 35 (2008) 2964 – 2987

[STE2003] Microarray Bioinformaticas
Dov Stekel
Cambridge University Press - 2003

[STO2000] DNA binding sites: representation and discovery
Gary D. Stormo
Bioinformatics - Vol. 16 no. 1 2000 Pages 16-23

[TAN2004] Biclustering Algorithms: A Survey
Amos Tanay, Roded Sharan y Ron Shamir
Mayo 2004

[YAN2012] Normalization: Bioconductor's marray package
http://www.bioconductor.org/packages/2.11/bioc/vignettes/marray/inst/doc/marrayNorm.pdf

5. APENDICES - Resultados completos

Apéndice A. Agrupamientos convencionales utilizando el algoritmo CLARA

Experimento GSE10391

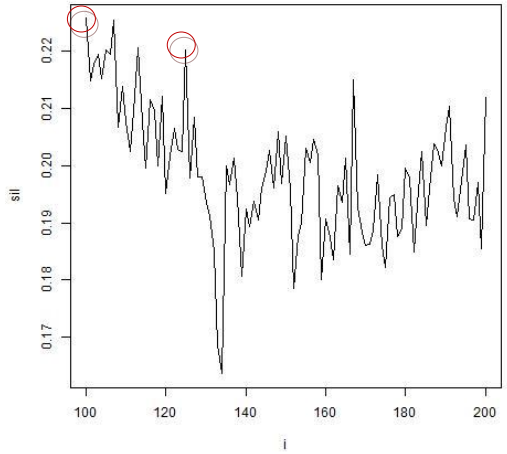
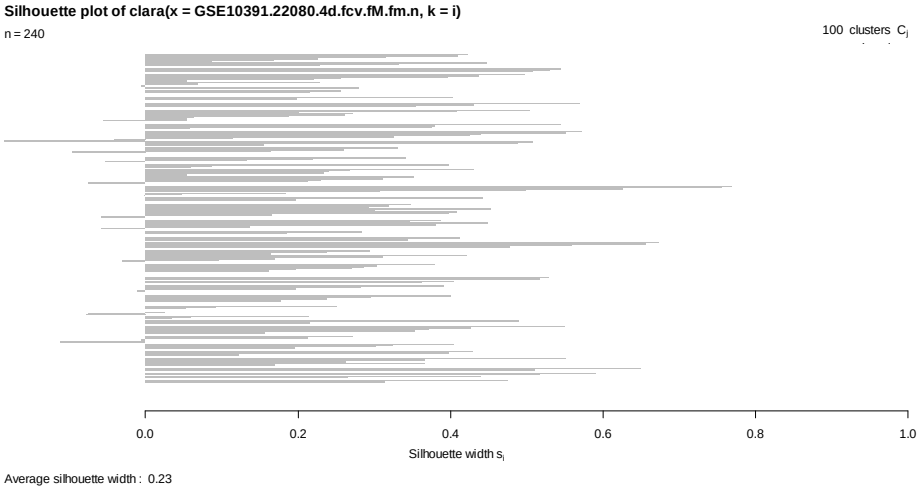


Figura Ap.1. Silhoutte width en función de la cantidad k de particiones



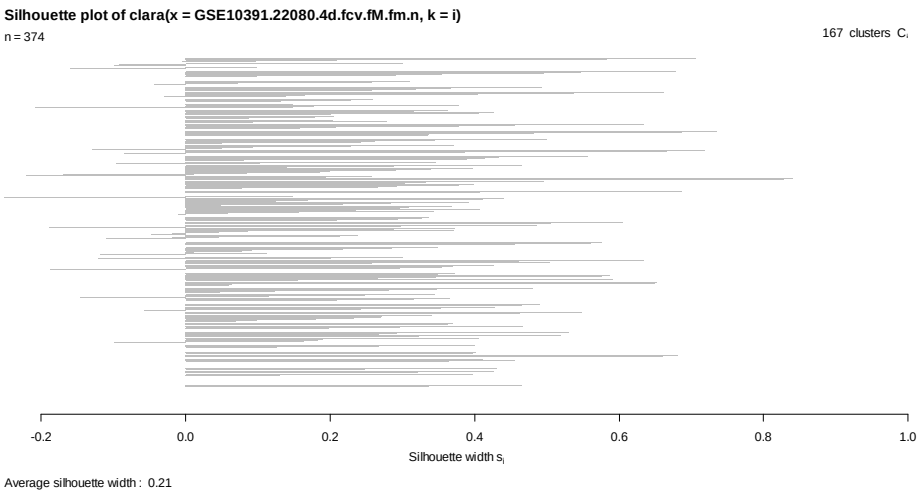


Figura Ap.2. Silhoutte para k=100 y k=157

Experimento GSE8639

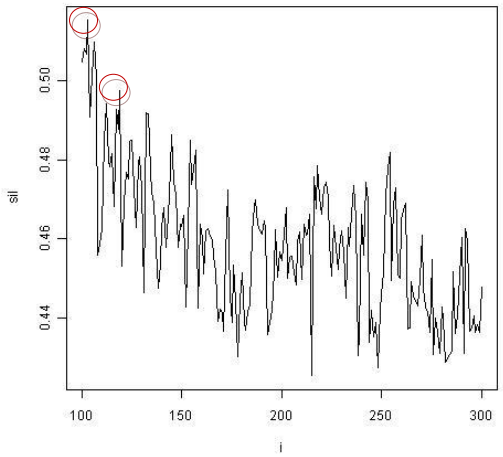
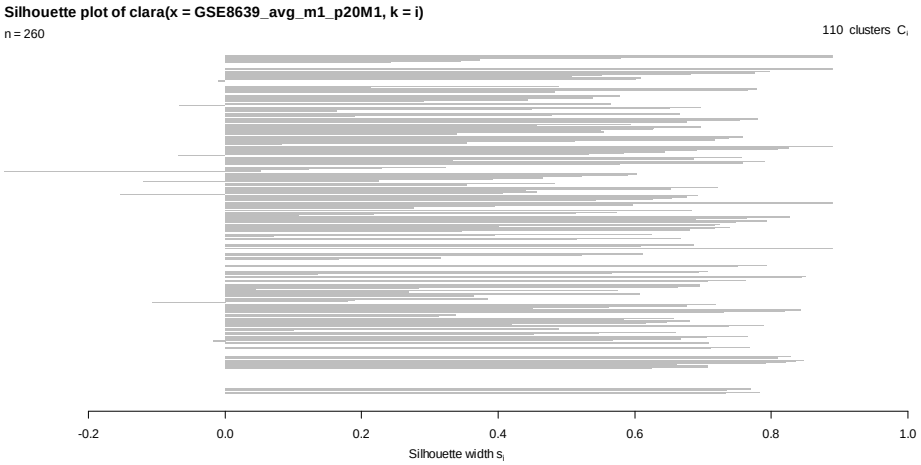


Figura Ap.3. Silhoutte width en función de la cantidad k de particiones



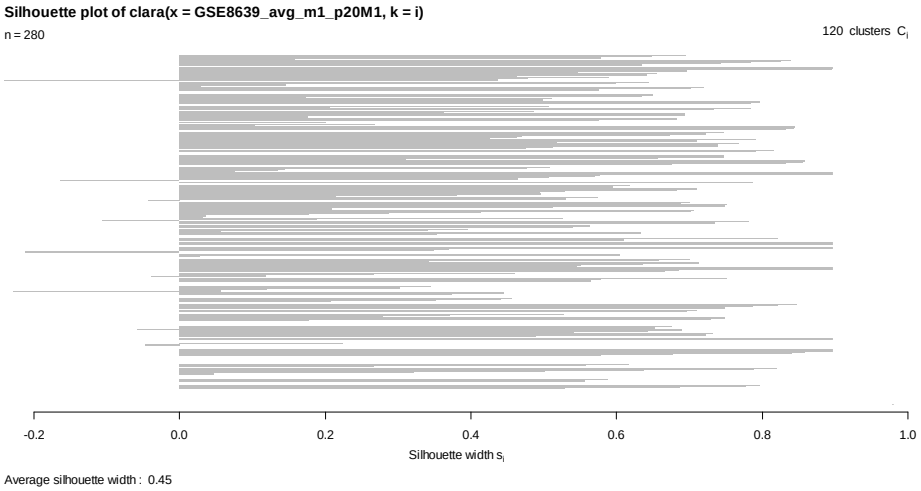


Figura Ap.4. Silhoutte para k=110 y k=120

Experimento GSE12364

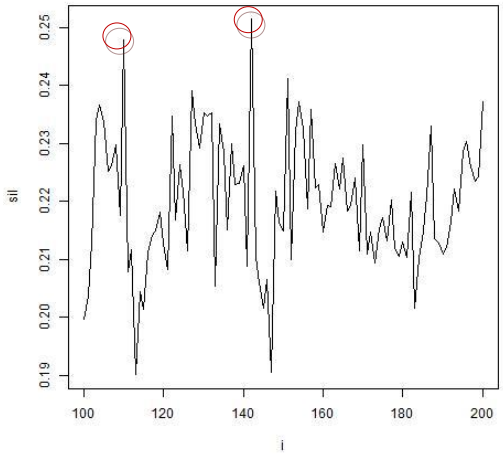
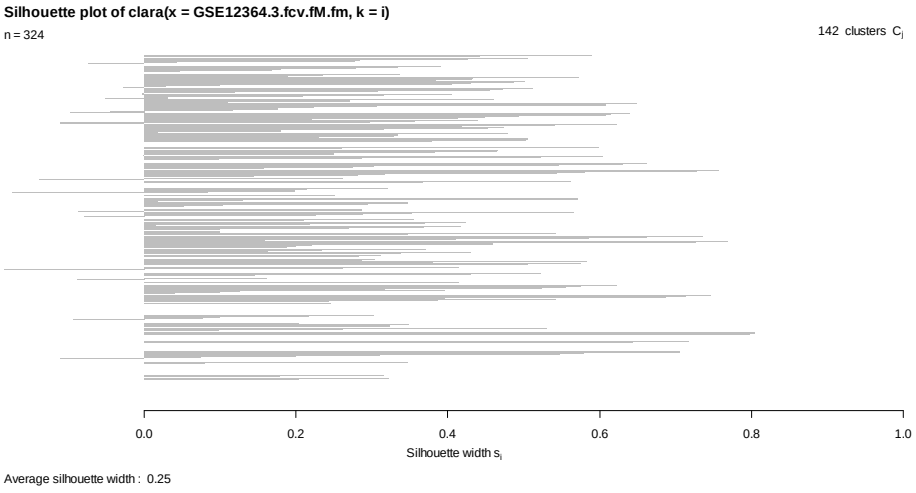


Figura Ap.5. Silhoutte para k=142 y k=183



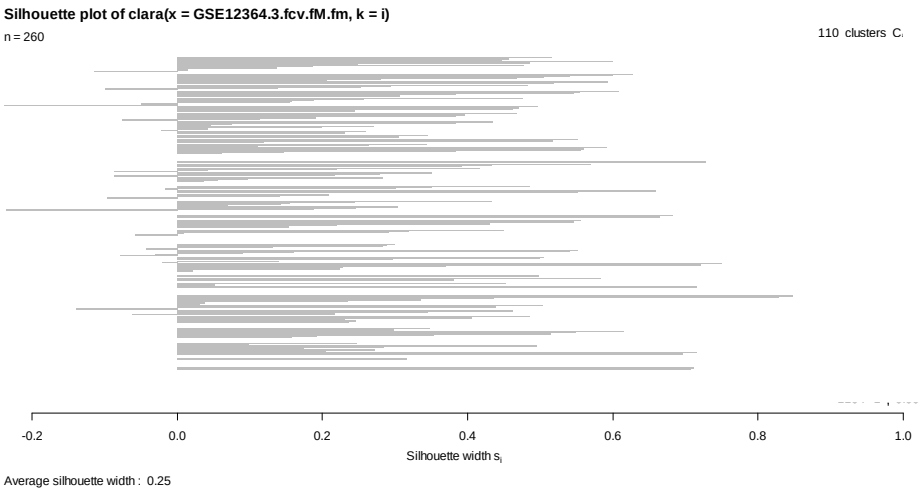


Figura Ap.6. Silhoutte para k=142 y k=183

Apéndice B. Agrupamientos convencionales utilizando el algoritmo HOPACH

Experimento GSE10391

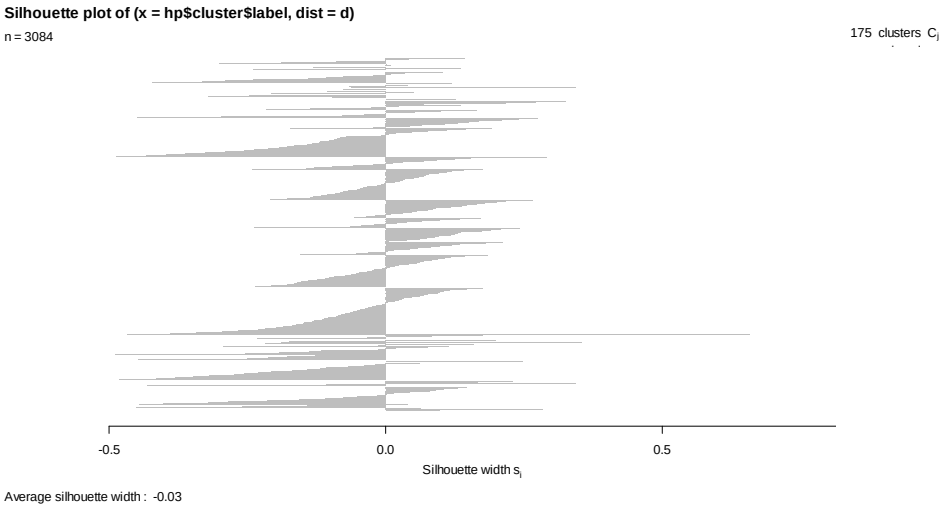


Figura Ap.7. Silhoutte (175 grupos determinados por HOPACH)

Experimento GSE8639

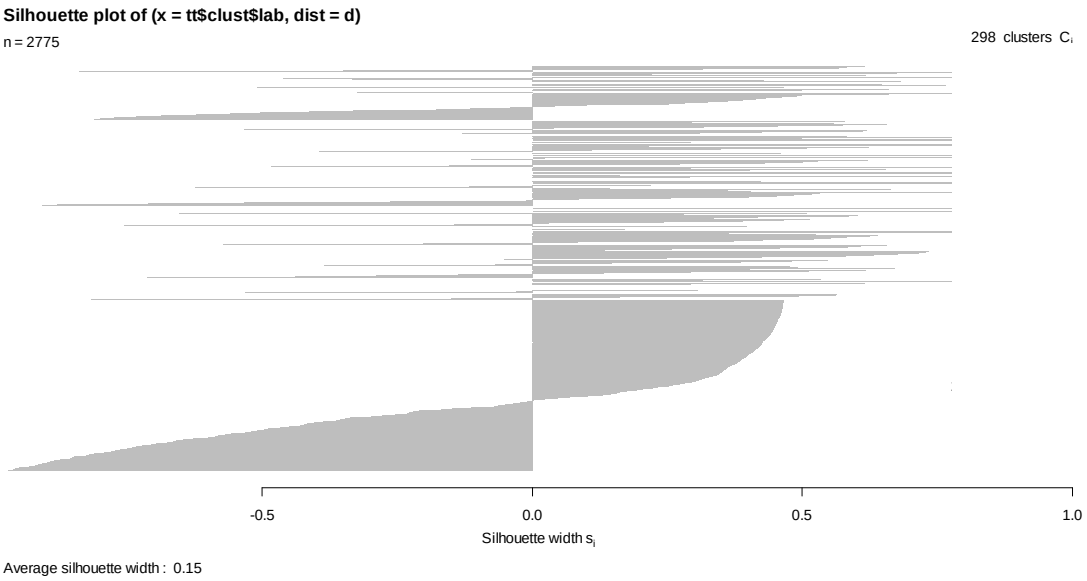


Figura Ap.8. Silhoutte (298 grupos determinados por HOPACH)

Experimento GSE12364

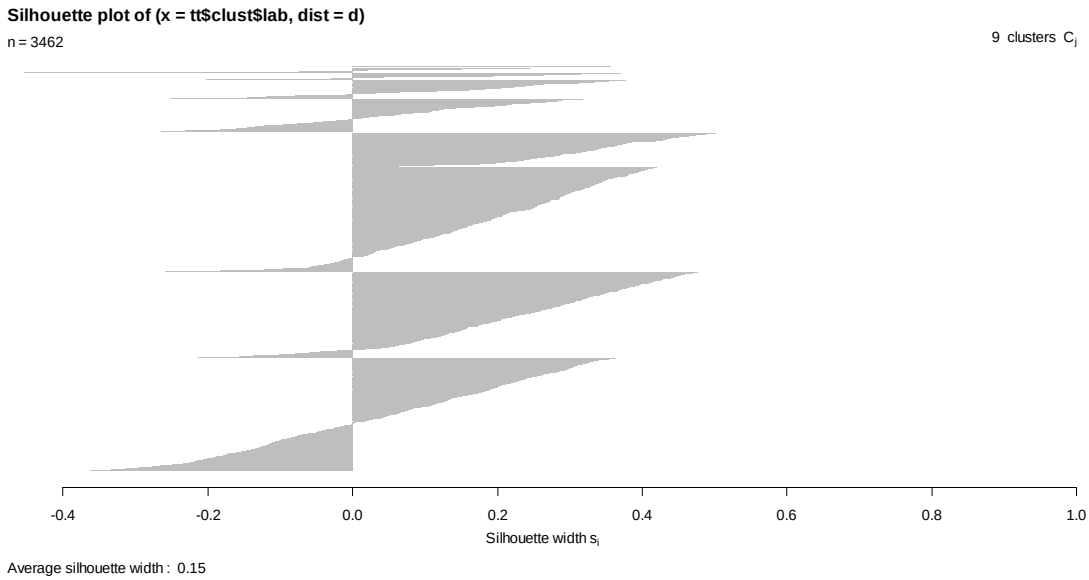


Figura Ap.9. Silhoutte (9 grupos determinados por HOPACH)

Apéndice C. Agrupamientos con filtros ANOVA utilizando el algoritmo CLARA

Experimento GSE6209

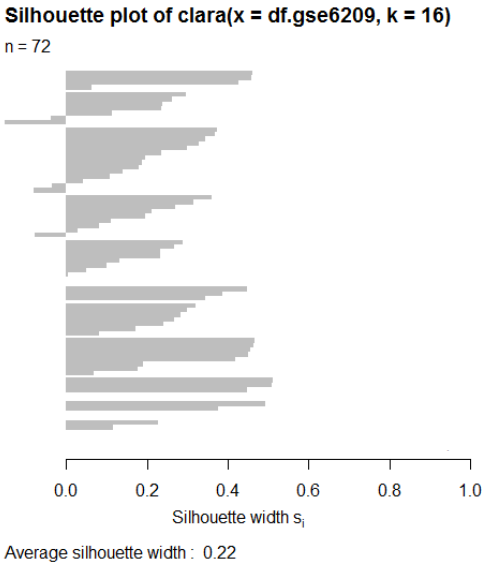


Figura Ap.10. Silhoutte para k=16

Experimento GSE12364

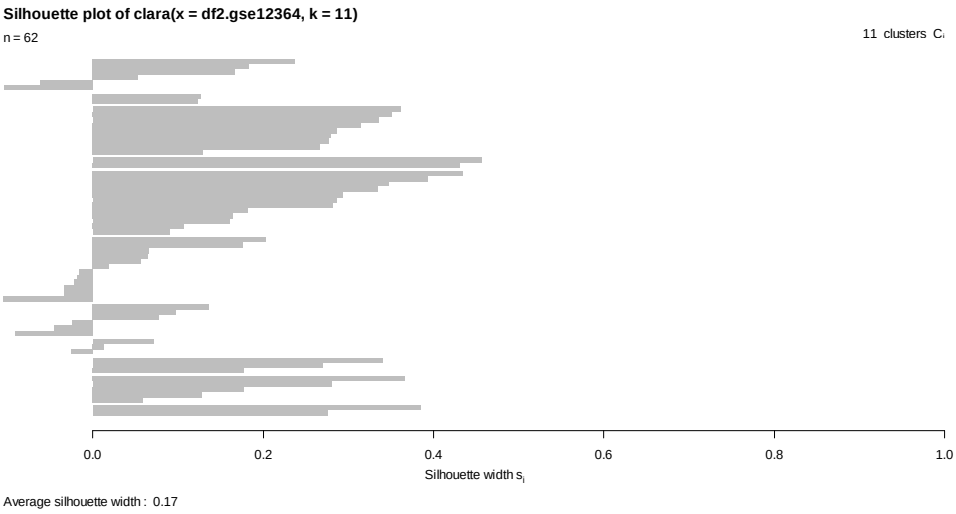


Figura Ap11. Silhoutte para k=11

Experimento GSE15976

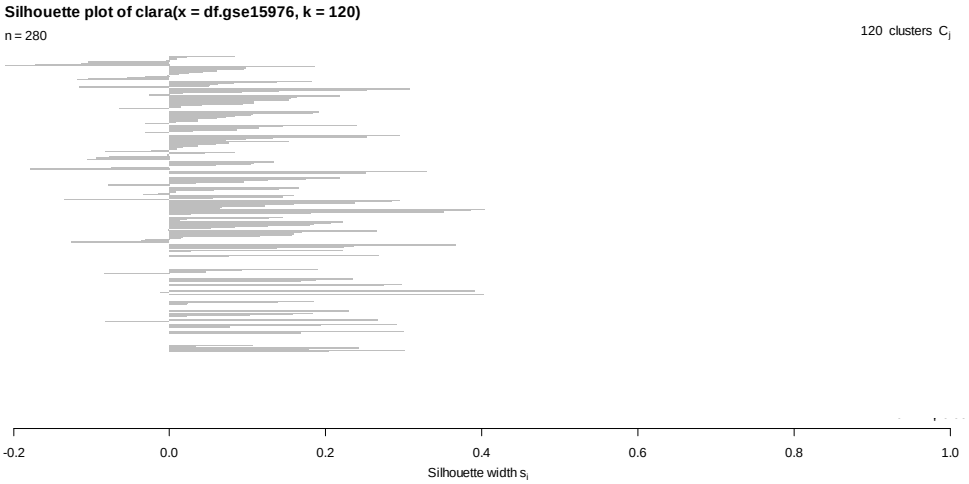


Figura Ap.12. Silhoutte para k=120

Experimento GSE9776

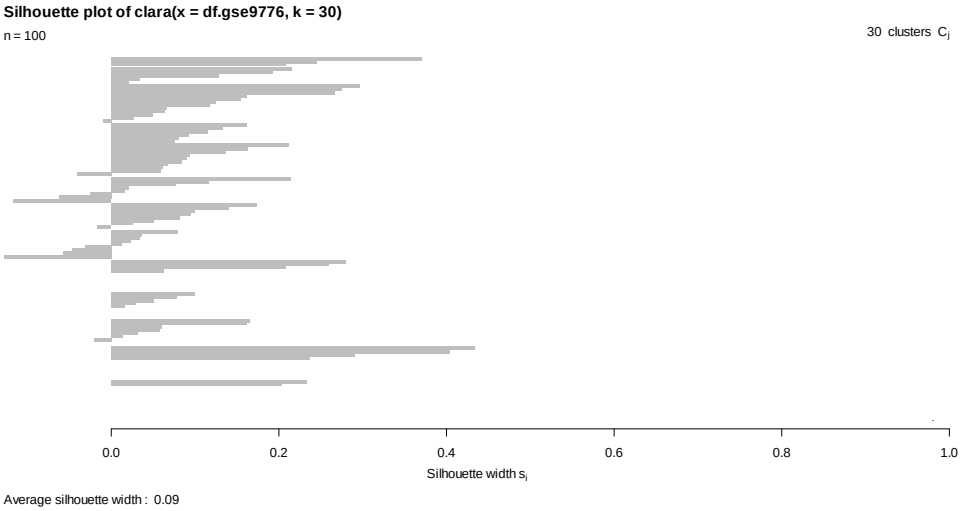


Figura Ap.13. Silhoutte para k=30

Experimento GSE365

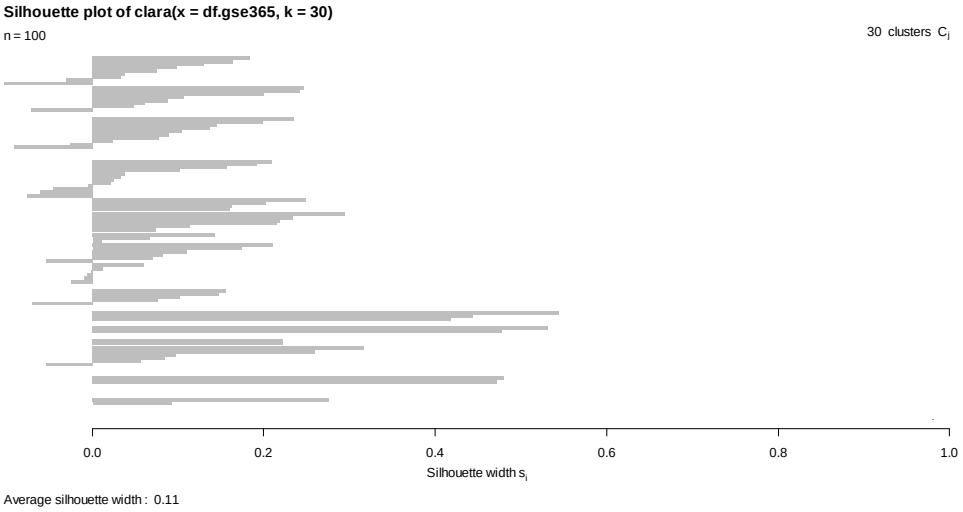


Figura Ap.14. Silhoutte para k=30

Experimento GSE7962

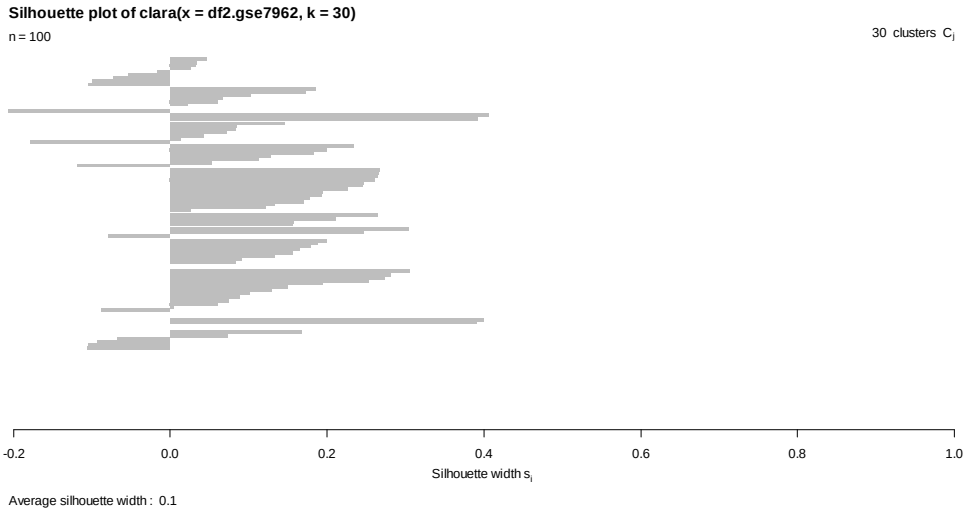


Figura Ap.15. Silhoutte para k=30

Apéndice D. Agrupamientos convencionales con filtro ANOVA aplicando el algoritmo HOPACH

Experimento GSE6209

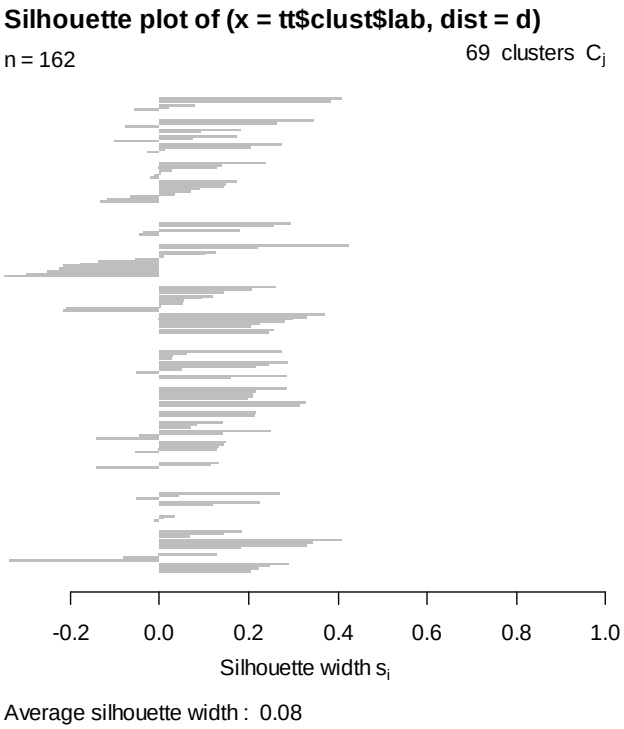


Figura Ap.16. Silhoutte para k=69

Experimento GSE12364

En este caso una vez aplicado el filtro el conjunto resultante queda con filas donde más del 50% de los valores están ausentes, motivo por el cual no es posible imputar valores mediante knn. Se procedió a eliminar dichas filas y continuar como en el resto de los experimentos. Se muestran los resultados sin sacar los valores ausentes y sacando los valores ausentes.

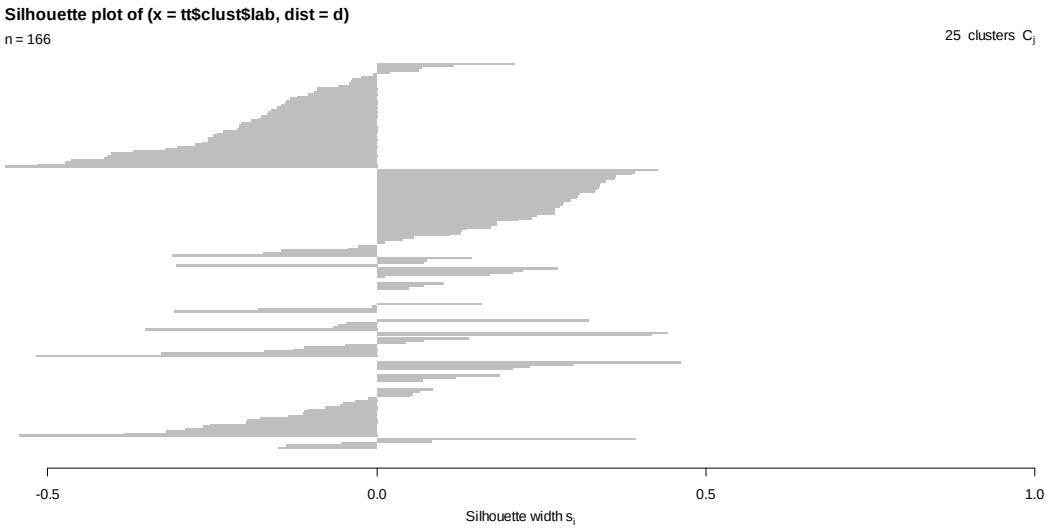


Figura Ap.17. Silhoutte con valores nulos para k=25

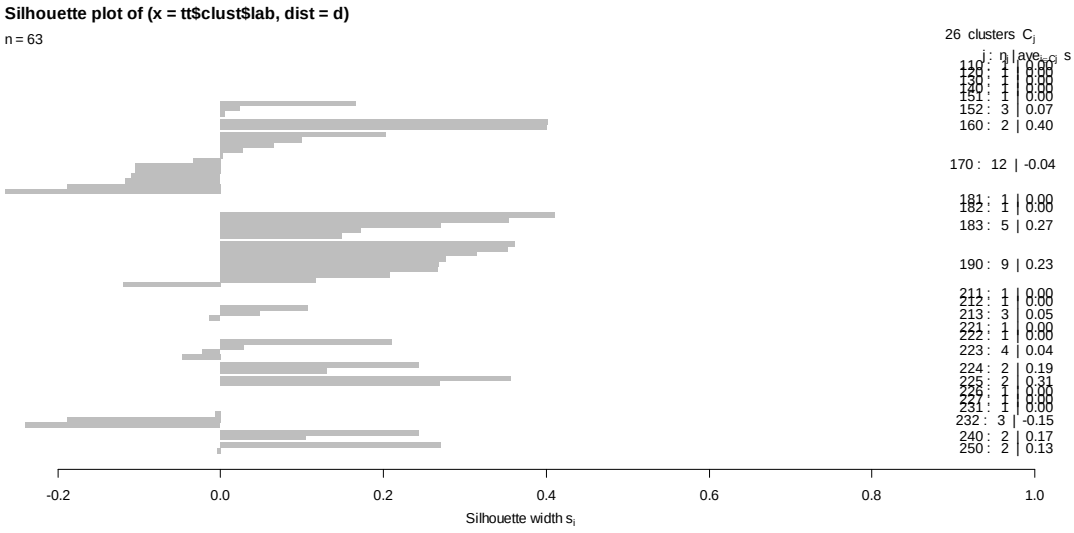


Figura Ap.18. Silhoutte sin valores nulos para k=26

Experimento GSE9776

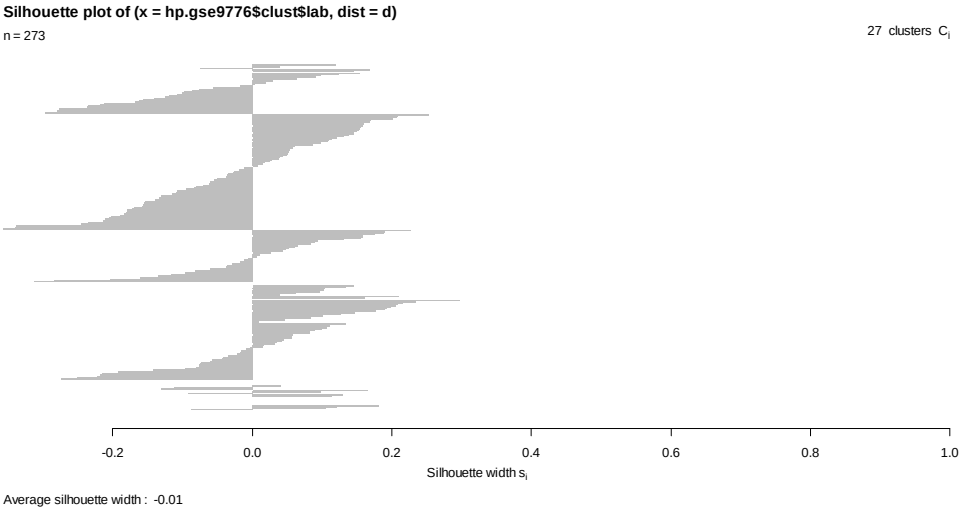


Figura Ap.19. Silhoutte para k=27

Experimento GSE365

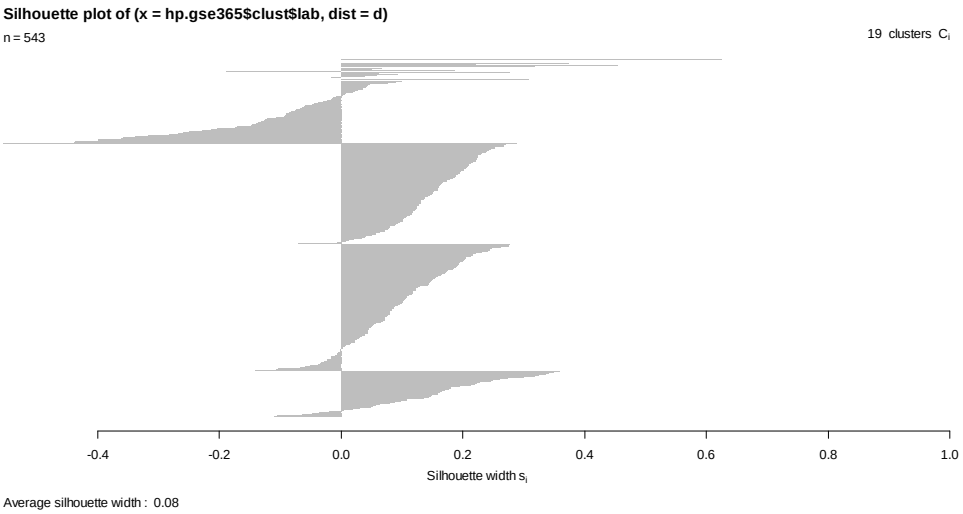


Figura Ap.20. Silhoutte para k=19

Experimento GSE7962

En este caso una vez aplicado el filtro el conjunto resultante queda con filas donde más del 50% de los valores están ausentes, motivo por el cual no es posible imputar valores mediante *knn*. Se procedió a eliminar dichas filas y continuar como en el resto de los experimentos. Se muestran los resultados sin sacar los valores ausentes y sacando los valores ausentes.

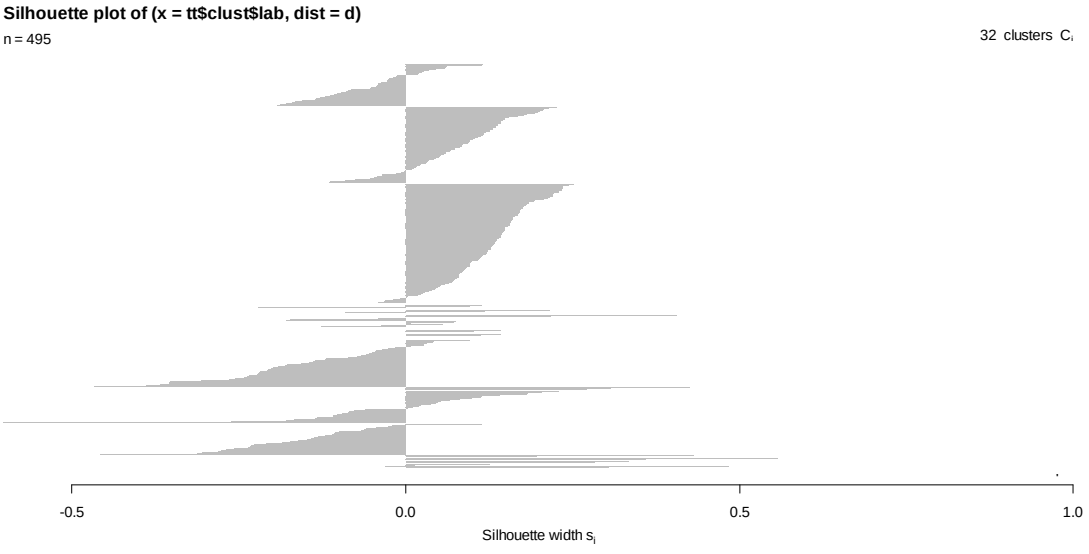


Figura Ap.21. Silhoutte con valores nulos para k=32

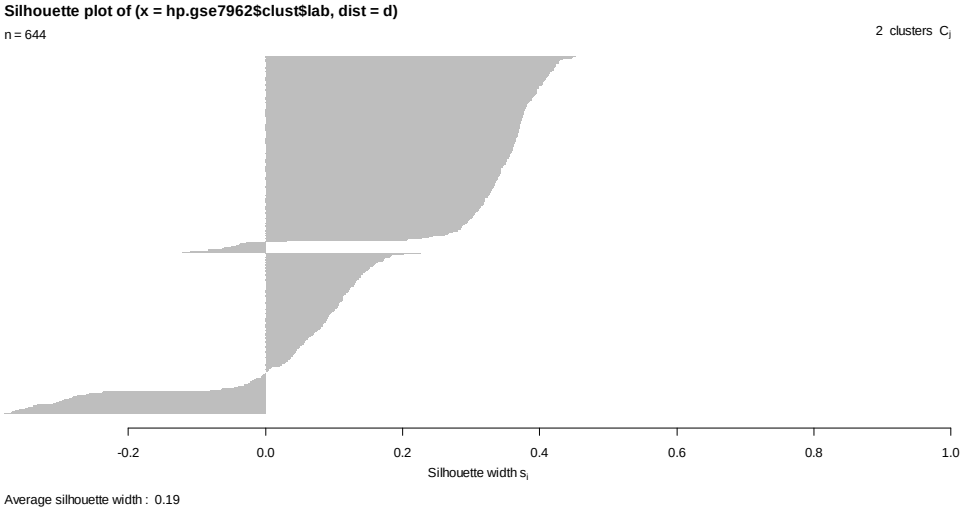


Figura Ap.22. Silhoutte sin valores nulos para k=2

Apéndice E. Biagrupamientos

Experimento GSE12364

Para este experimento el método BCC no encuentra ningún biagrupamiento.

Por otro lado, el método Questmet encuentra 2 biagrupamientos, donde el segundo tiene una cantidad interesante de filas y de columnas, y buenas medidas para valores constantes y coherencia aditiva.

Bicluster Questmet para el experimento GSE12364						
Identificador	**# Fil.**	**# Col.**	**Validación**			
			Const.	**Adit.**	**Mult.**	**Signo**
BC1	**42**	10	0,66	0,81	22.71	2,43
BC2	**14**	12	0,74	0,65	11,01	1,97

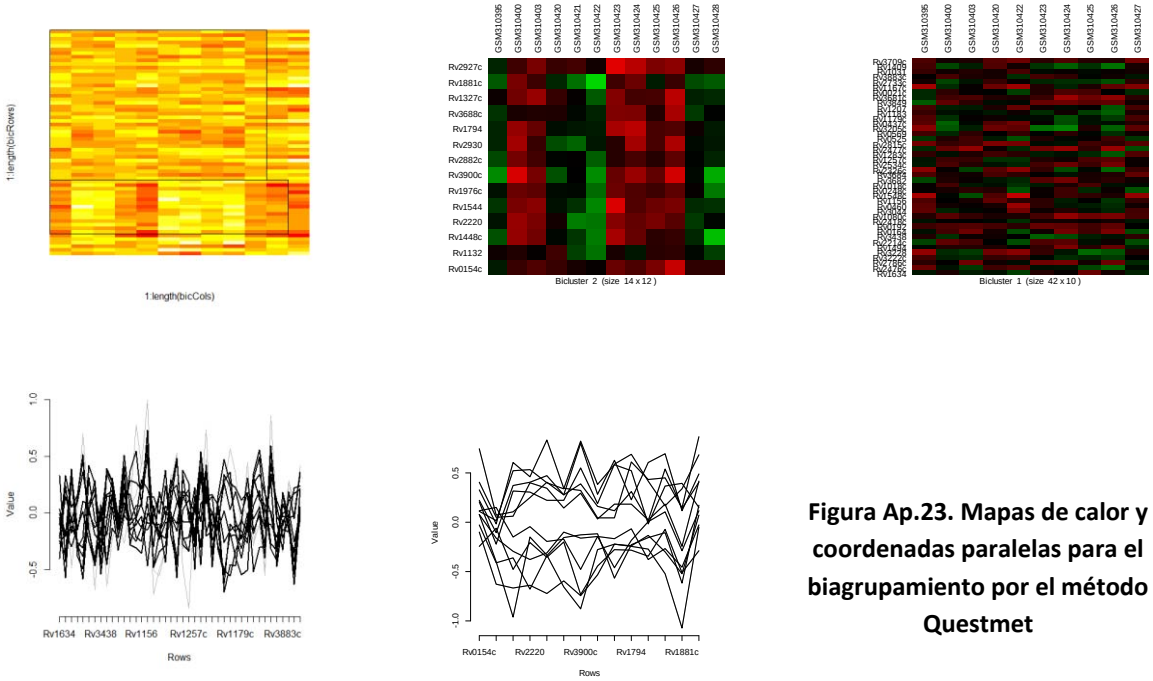


Figura Ap.23. Mapas de calor y coordenadas paralelas para el biagrupamiento por el método Questmet

Experimento GSE9776

Par este experimento el método BCC encuentra solamente el biagruoamiento 1ue incluye a todas las filas y a todas las columnas, por lo tanto se descarta.

Para el caso del método Questmet, encuentra 5 biagrupamientos, con buenas medidas de valores constantes y coherencia aditiva. En particular los correspondientes al BC3 y BC4 tienen además una cantidad de genes adecuada y cubren muestras que abarcan réplicas de distintas condiciones. Por ejemplo, si se observan las condiciones del experimento en la Fig. 41, y las comparamos con las condiciones cubiertas por el grupo 4 marcadas en color rojo (GSM241392, GSM241433, GSM241434, GSM241437, GSM241439, GSM241441, GSM241442, GSM241443), vemos que están presentes 3 condiciones distintas y para todas sus réplicas.

Bicluster Questmet para el experimento GSE9776						
Identificador	# Fil.	# Col.	Validación			
			Const.	Adit.	Mult.	Signo
BC1	**140**	8	0,29	0,30	6435,21	2.07
BC2	**85**	5	0,20	0,16	4560.44	1,51
BC3	**22**	6	0,26	0,25	10,06	1,65
BC4	**9**	8	0,17	0,15	2338,79	1,69
BC5	**7**	9	0,30	0,26	122,89	1,53

Muestras del experimento GSE6209	
Muestra	Descripción
GSM241391	Log phase 2 horas 1ug/ml INH tratado réplica 1
GSM241392	Log phase 6 horas 1ug/ml INH tratado réplica 1
GSM241431	Log phase 2 horas 1ug/ml INH tratado réplica 2
GSM241432	Log phase 2 horas 1ug/ml INH tratado réplica 3
GSM241433	Log phase 6 horas 1ug/ml INH tratado réplica 2
GSM241434	Log phase 6 horas 1ug/ml INH tratado réplica 3
GSM241437	Log phase 2 horas 1ug/ml INH treated KatG knockout replicate 1
GSM241439	Log phase 2 horas 1ug/ml INH treated KatG knockout replicate 2
GSM241441	Log phase 2 horas 1ug/ml INH treated KatG knockout replicate 3
GSM241442	Nutrientes restringidos 2 horas 1ug/ml INH tratado réplica 1
GSM241443	Nutrientes restringidos 2 horas 1ug/ml INH tratado réplica 2
GSM241450	Oxígeno agotado 6 horas 1ug/ml INH tratado réplica 1
GSM241662	Oxígeno agotado 6 horas 1ug/ml INH tratado réplica 3
GSM241663	Fibra hueca encapsulado del ratón (Mouse hollow fiber encapsulated) 25mg/kg isoniacida 3 dose 6 horas réplica 1
GSM241664	Fibra hueca encapsulado del ratón 25mg/kg isoniacida 3 dose 6 horas réplica 2
GSM241665	Fibra hueca encapsulado del ratón 25mg/kg isoniacida 3 dose 6 horas réplica 3

Figura Ap.24. Condiciones para el experimento GSE9776

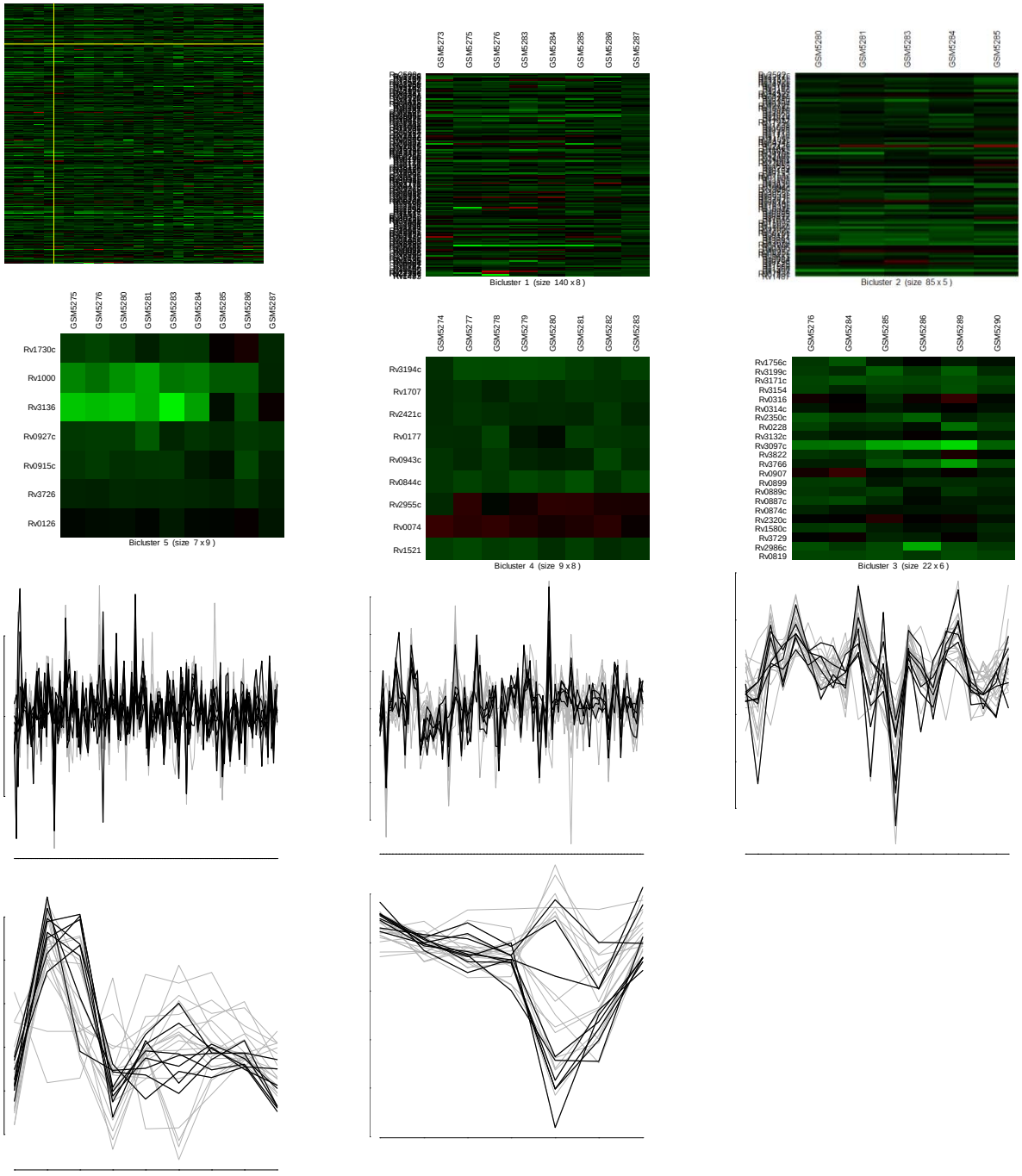


Figura Ap.25. Mapas de calor y coordenadas paralelas para el biagrupamiento por el método Questmet

Experimento GSE365

Finalmente veamos los resultados de aplicar el biagrupamiento Quetmet al experimento GSE365, donde podemos considerar el segundo grupo por poseer una cantidad adecuada de filas y de columnas, cubriendo varias condiciones, y con buenas medidas de valores constantes y coherencia aditiva.

Bicluster Questmet para el experimento GSE365						
Identificador	# Fil.	# Col.	Validación			
			Const.	Adit.	Mult.	Signo
BC1	**526**	26	0,47	0,47	12645.71	3,92
BC2	**11**	22	0,50	0,61	22,38	2,27

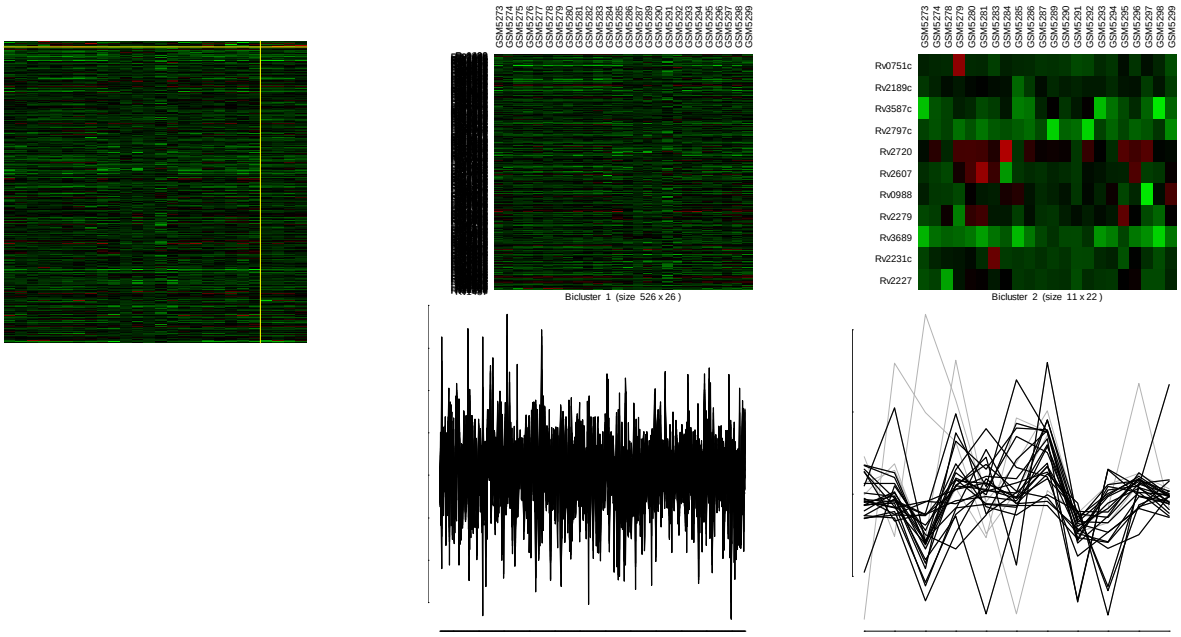


Figura Ap.26. Mapas de calor y coordenadas paralelas para el biagrupamiento por el método Questmet

Funciones principales para el análisis de datos provenientes de experimentos con microarreglos

Autor: Guillermo Henrión

Fecha: Noviembre 2013

A continuación se detallan las funciones principales implementadas en el lenguaje R, correspondiente al flujo de ejecución citado en el desarrollo de la tesis.

getGSE, getGDS

Descripción

Función que recupera los datos de un experimento, dado como parámetro, desde el NCBI dado

Uso

getGSE <- function(mb_dataset)

getGDS <- function(mb_dataset)

Argumentos

mb_dataset Nombre del experimento a recuperar

cluster.semantics

Descripción

Una función para obtener información semántica y ontológica para un agrupamiento desde un objeto de análisis silhouette.

Uso

cluster.semantics <- function(silClus)

Argumentos

silClus Objeto silhoutte de un agrupamiento

TOSimGenes

Descripción

Una función para calcular el Term Overlap de dos genes, dada la lista de sus anotaciones.

Uso

TOSimGenes <- function (annotSet1, annotSet2)

Argumentos

annotSet1, annotSet2 Conjunto de anotaciones para cada uno de los genes

TOSim_XXnorm_stats

Descripción

Funciones para calcular el Term Overlap dada una lista de genes (para las tres ontologías BP, CC y MF).

Uso

TOSim_BPnorm_stats <- function (gs)

TOSim_CCnorm_stats <- function (gs)

TOSim_MFnorm_stats <- function (gs)

Argumentos

gs Lista de genes

Filtros

Descripción

Funciones para aplicar filtros

Uso

filtrarXcv <- function(dataset, a=0.5, b=Inf)

excluirGenesMenores <- function(dataset, p=1, A=0.1)

excluirGenesMayores <- function(dataset, p=1, A=1)

Argumentos

dataset	Conjunto de datos a filtrar
a	Cota inferior
b	Cota superior
p	Proporción de muestras que cumplen la condición
A	Valor por arriba (o debajo) del cual el gen es excluido

cluster.info

Descripción

Una función para obtener información ontológica sobre un agrupamiento

Uso

cluster.info <- function(clus, x)

Argumentos

clus	Agrupamiento

x Miembro dentro del agrupamiento

biclusters

Descripción

Una función para obtener los biagrupamientos

Uso

biclust <- function()

generaPam

Descripción

Una función para obtener los agrupamientos mediante Pam

Uso

generarPam<- function()

generaClara

Descripción

Una función para obtener los agrupamientos mediante Clara

Uso

generarClara<- function()

generaHopach

Descripción

Una función para obtener los agrupamientos mediante Hopach

Uso

generarHopach<- function()